The Dissertation Committee for Wenhao Wang
certifies that this is the approved version of the following dissertation:

# An Algorithm of a Fully Conservative Volume Corrected Characteristics-Mixed Method for Transport Problems

Committee:

Todd Arbogast, Supervisor

Clint Dawson

Chieh-Sen (Jason) Huang

Yen-Hsi Richard Tsai

Mary Wheeler

# An Algorithm of a Fully Conservative Volume Corrected Characteristics-Mixed Method for Transport Problems

**by**

**Wenhao Wang, B.S., M.S.**

**DISSERTATION**

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

**DOCTOR OF PHILOSOPHY**

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2009

For everyone who
helped and supported me along the way,
THANKS Y'ALL.

# Acknowledgments

I give my sincerest gratitude to my advisor, Todd Arbogast, a gracious mentor for his patience, encouragement and understanding. I am grateful to him for his guidance and support in getting my graduate career started on the right foot and providing me with the foundation for becoming a professional researcher. I would also like to thank the Institute for Computational Engineering and Sciences at the University of Texas at Austin, especially those members of my doctoral committee for their input, effort, and valuable discussions.

I am grateful to my mother, who gives me constant understanding, encouragement and support, both physically and spiritually. During all these years when I studied abroad, there is no doubt that I could not have made the achievement today without her selfless love and caring.

Finally, I am thankful to all of the friends and colleagues along the way who helped and supported me these years. For everything you have done for me, thank you all.

# An Algorithm of a Fully Conservative Volume Corrected Characteristics-Mixed Method for Transport Problems

Publication No. _____

Wenhao Wang, Ph.D.
The University of Texas at Austin, 2009

Supervisor: Todd Arbogast

A basic phenomenon modeled computationally is tracer transport in a flow field, such as in porous medium simulation. We analyze the stability and convergence of a fully conservative characteristic method, the Volume Corrected Characteristics-Mixed Method [4] (VCCMM) applied to advection of a dilute tracer in an incompressible flow. Numerical tests for the optimal convergence rate match the results of our theoretical proof. We avoid the CFL constraint on the time step size and obtain a higher order convergence rate compared with Godunov's method. We describe the implementation of the VCCMM, where we feature and define a polyline class for the volume computation of trace-back regions. Some numerical examples show that large time steps can be used in practice, no overshoot or undershoot arises in the solution, and less numerical diffusion is produced compared with Godunov's method. An application to a nuclear waste disposal problem is also presented, where we simulate the processes of advection, reaction, and diffusion of radioactive elements in a simplified far field model. Finally, an extension of the VCCMM is developed for compressible flows, and a stability and convergence analysis is presented.

# Table of Contents

# List of Symbols

## English Symbols

| | |
|---|---|
| $\mathbb{R}^d$ | Euclidean space of dimension $d$ ..... 1 |
| $\mathbb{R}^{m\times n}$ | space of real matrices with $m$ rows and $n$ columns ..... 15 |
| $R$ | ring for adjustment ..... 39 |
| $\check{R}^n$ | exact trace-back ring of $R$ at time $t^n$ ..... 39 |
| $\tilde{R}^n$ | adjusted ring of $R$ at time $t^n$ ..... 39 |
| $\mathcal{S}$ | space boundary of a space-time region ..... 11 |
| $T_{\mathbf{y}}$ | translation operator in direction $\mathbf{y}$ ..... 26 |
| $\mathcal{T}_h$ | collection of elements in a mesh of domain ..... 10 |
| $\mathcal{T}_{h,P}$ | collection of elements in mesh $\mathcal{T}_{h,P}$ which represent locations of production wells ..... 10 |
| $\mathbf{u}$ | velocity field ..... 1 |
| $\tilde{\mathbf{u}}$ | perturbed velocity field ..... 18 |
| $\mathbf{v}$ | interstitial velocity field ..... 10 |
| $\tilde{\mathbf{v}}$ | perturbed interstitial velocity field ..... 89 |
| $\mathbf{V}$ | vector space in a pair of mixed spaces ..... 50 |
| $\mathbf{V}_h$ | vector space in a pair of mixed finite element spaces ..... 51 |
| $W$ | scalar space in a pair of mixed spaces ..... 50 |
| $W_h$ | scalar space in a pair of mixed finite element spaces ..... 51 |
| $W_h(\Omega)$ | space of piecewise constant functions on the mesh ..... 51 |
| $\check{\mathbf{x}}$ | trace-back characteristic ..... 6 |
| $\hat{\mathbf{x}}$ | trace-forward characteristic ..... 6 |
| $\tilde{\mathbf{x}}$ | perturbed trace-back characteristic ..... 18 |

**Greek Symbols**

| | |
|---|---|
| $\Delta t$ | maximal time step ..... 6 |
| $\Delta t^n$ | time step of interval $J^n$ ..... 27 |
| $\Delta t_{\text{CFL}}$ | CFL restricted time step ..... 63 |
| $\varepsilon_{\text{flow}}$ | $L^\infty$-error of flow approximation ..... 85 |
| $\varepsilon_{\text{tol}}$ | relative tolerance error of volumes in trace-back points adjustment 57 |

**Other Symbols**

# List of Tables

# List of Figures

# Chapter 1

# Introduction

A basic phenomenon modeled computationally is tracer transport in a flow field, as might arise in a porous medium or a shallow water or atmospheric system. We concentrate on the problem as it arises in porous media simulation, though many of the ideas carry over to other contexts. Thus we envision, e.g., transport of a tracer or contaminant in the groundwater, or the transport of immiscible phases in an oil/water petroleum reservoir system.

A porous medium or a porous material is a solid, which is often called the matrix, permeated by an interconnected network of pores filled with liquid or gas phases. Usually both the solid matrix and the pore network are assumed to be continuous. Many natural substances, such as rocks, soils, and sand, can be considered as porous media. For example, a petroleum reservoir is a porous medium that contains hydrocarbons. A porous medium is characterized by its porosity, permeability, and the properties of its constituents. In mathematical terminology, a porous medium is the closure of a subset $\Omega$ of the Euclidean space $\mathbb{R}^d$ ($d = 1, 2$ or $3$).

## 1.1 Problem

In this dissertation, we will mainly study the simplest case of a linear, incompressible tracer transport problem in a porous medium. The bulk or ambient fluid flows with a velocity $\mathbf{u}(\mathbf{x}, t)$ satisfying the incompressibility condition

$$\nabla \cdot \mathbf{u} = q,$$

where $q(\mathbf{x}, t)$ is a given external source or sink function, such as a well in a porous medium. A tracer solute species of concentration $c(\mathbf{x}, t)$ transports within the bulk fluid. Assuming that it does not change the overall velocity $\mathbf{u}$, the concentration will satisfy an advection-diffusion equation of the form

$$(\phi c)_t + \nabla \cdot (c\mathbf{u} - \mathbf{D}\nabla c) = q_c := c_I q^+ + c q^-, \tag{1.1}$$

where $\phi(\mathbf{x})$ is the storage factor of the medium called porosity, subscript $t$ is time partial differentiation, $\mathbf{D}(\mathbf{x}, t)$ is the diffusion-dispersion tensor (which may also depend on $\mathbf{u}$ and is assumed to be bounded and positive definite), $q^+(\mathbf{x}, t) = \max\{q, 0\} \geq 0$ is $q$ when $q > 0$ and $q^-(\mathbf{x}, t) = q - q^+ \leq 0$, and $c_I(\mathbf{x}, t)$ is the given concentration of injected fluid.

Discretization of (1.1) often uses an operator splitting technique [27] to isolate the hyperbolic and parabolic parts of the equation. That is, over a time step, one first approximates the hyperbolic part of the operator,

$$(\phi c)_t + \nabla \cdot (c\mathbf{u}) = q_c, \tag{1.2}$$

and then the parabolic part,

$$(\phi c)_t - \nabla \cdot (\mathbf{D}\nabla c) = 0. \tag{1.3}$$

The hyperbolic part (1.2), which models pure transport of the fluid particles, is the most delicate to approximate well, and we only study the approximation of this part in the dissertation. The rest of the operator, $-\nabla \cdot (\mathbf{D}\nabla c)$, is diffusive, and many excellent techniques are known for (1.3), including finite elements [12], mixed methods [13, 47], and discontinuous Galerkin methods [8].

Fixed grid methods, such as Godunov's method, the MUSCL scheme, Lax-Wendroff, ENO and WENO methods, etc., have enjoyed a great deal of success, as they can be made to be reasonably accurate while minimizing numerical diffusion and dispersion [10, 41]. However, they are subject to a Courant-Friedrichs-Lewy

(CFL) constraint to maintain stability, severely restricting the time step size. Often this restriction requires that a very large number of time steps be taken, resulting in significant numerical diffusion/dispersion and smearing of otherwise sharp fronts over time. Moreover, Godunov's method, e.g., does not naturally extend to multiple space dimensions, but rather is usually applied as an essentially one-dimensional method normal to each grid element face [25].

## 1.2 Outline

The outline of the rest of the dissertation follows. In Chapter 2, we give a background of development of characteristic methods, where a review of conservative characteristic methods and a derivation of local mass constraints are presented followed by a brief description of the Volume Correction Algorithm for the *Volume Corrected Characteristics-Mixed Method* (VCCMM). Also, a stability analysis of the VCCMM is provided.

Chapter 3 gives a convergence analysis of the VCCMM, wherein an $L^1$-error estimate is derived. The technique we employ is to introduce an approximation of the $L^1$-error and use a Kuznetsov type proof [40]. The major difficulty of the convergence proof is to construct and estimate the error of the perturbed velocity field.

In Chapter 4, we describe the implementation of the VCCMM, which includes the flow approximation and transport approximation. In particular, to compute volumes of trace-back regions approximated by polygons, we define and use a polyline class in the code.

In Chapter 5, we give the results of some numerical tests of a rotating pollutant problem and a quarter of five-spot pattern problem in petroleum reservoir simulations. We also test the optimal convergence rate of the VCCMM and make some comparisons with the Characteristics-Mixed Method and the first order Go-

dunov's method.

Chapter 6 describes an application to a nuclear waste disposal problem and gives results of numerical simulations, where we simulate all the processes of advection, reaction, and diffusion of radioactive species.

In Chapter 7, we extend the VCCMM to the problem of compressible flows. We describe both the flow and transport approximations. Then we extend the stability analysis and convergence analysis of incompressible flows to this case.

Finally, we make some conclusions and lay out some directions for future research in Chapter 8.

A table of symbols can be found for easy reference just after the table of contents, and is followed by a list of tables and a list of figures. A bibliography and an index follow the last chapter.

# Chapter 2

# The Volume Corrected Characteristics-Mixed Method (VCCMM)

We consider the problem of incompressible dilute miscible tracer transport on a confined and bounded domain $\Omega \subset \mathbb{R}^d$. A dilute miscible tracer of concentration $c(\mathbf{x}, t)$ in an incompressible bulk fluid moving according to the velocity field $\mathbf{u}(\mathbf{x}, t)$ satisfies the advection system

$$\nabla \cdot \mathbf{u} = q \qquad \text{in } \Omega \times J, \tag{2.1}$$

$$(\phi c)_t + \nabla \cdot (c\mathbf{u}) = q_c \qquad \text{in } \Omega \times J, \tag{2.2}$$

$$\mathbf{u} \cdot \boldsymbol{\nu} = 0 \qquad \text{on } \partial\Omega \times J, \tag{2.3}$$

$$c(\mathbf{x}, 0) = c^0(\mathbf{x}) \qquad \text{in } \Omega, \tag{2.4}$$

where $J = [0, \infty)$ is the time interval, porosity $\phi = \phi(\mathbf{x}) \in [\phi_*, 1]$ with some constant $\phi_* > 0$, $c^0 = c^0(\mathbf{x})$ is the initial concentration, and $\boldsymbol{\nu}$ is the outward unit normal vector with respect to $\partial\Omega$. The meaning of a *dilute* tracer is that we assume $c$ does not change the overall velocity $\mathbf{u}$.

## 2.1 Background of Characteristic Methods

Equation (2.2) is linear, so it admits only contact discontinuities: shocks and rarefactions do not form. Thus characteristics (wave velocity paths) and streamlines (particle velocity paths) coincide. In the absence of diffusion, fluid particles simply travel along the characteristics or streamlines of the equation,

defined by $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{x}, t)$ satisfying the ordinary differential equation

$$\hat{\mathbf{x}}_t = \frac{\mathbf{u}(\hat{\mathbf{x}}, t)}{\phi(\hat{\mathbf{x}})}, \qquad t \in J,$$
$$\hat{\mathbf{x}}(0) = \mathbf{x}.$$

Moving mesh and characteristic methods have been developed to exploit this observation and thereby avoid any CFL constraint. CIR method [22] was proposed by Courant, Isaacson and Rees in 1952 by tracing points back along characteristics and evaluating characteristic variables based on interpolation of near grid values. Characteristic methods became viable in 1982 when Douglas and Russell introduced a Lagrangian formulation called the Modified Method of Characteristics (MMOC) [28, 32, 33] (see also [44]). In their method, one traces along the characteristics *backward* in time over the time step $J^n := [t^n, t^{n+1})$, giving $\check{\mathbf{x}}(\mathbf{x}, t)$ defined by

$$\check{\mathbf{x}}_t = \frac{\mathbf{u}(\check{\mathbf{x}}, t)}{\phi(\check{\mathbf{x}})}, \quad t \in J^n, \tag{2.5}$$
$$\check{\mathbf{x}}(t^{n+1}) = \mathbf{x}. \tag{2.6}$$

One approximates the characteristic derivative by a finite difference in the characteristic direction; that is, at each grid point $\mathbf{x}$,

$$\frac{dc}{dt}(\check{\mathbf{x}}(\mathbf{x}, t), t) = c_t(\check{\mathbf{x}}, t) + \frac{\mathbf{u}(\check{\mathbf{x}}, t)}{\phi} \cdot \nabla c(\check{\mathbf{x}}, t) \approx \frac{c(\mathbf{x}, t^{n+1}) - c(\check{\mathbf{x}}(\mathbf{x}, t^n), t^n)}{\Delta t}.$$

The approximation of (2.2) is then

$$\phi \frac{c(\mathbf{x}, t^{n+1}) - c(\check{\mathbf{x}}(\mathbf{x}, t^n), t^n)}{\Delta t} = (c_I - c) q_+.$$

However, the fluid must obey two physical principles: (1) *tracer mass conservation* and (2) mass conservation of the incompressible bulk fluid, or loosely speaking, by incompressibility, *volume conservation*. Numerical methods should respect both these conservation principles over the computational mesh (i.e., locally). We call such methods *fully conservative*. The first principle is well known,

and the second was emphasized by Arbogast and Huang in a recent paper [4]. To see this, we need merely consider the conservation principle for the rest of the fluid, which is an equation like (2.2) for the ambient fluid concentration $1 - c$. The sum of this equation and (2.2) is (2.1). That is, the volume constraint is implicit in (2.1) and holds on the differential level; however, it needs not hold on the discrete level.

Because MMOC is based on points, it violates *both* local mass and volume constraints. A modification of the method, MMOC with Adjusted Advection (MMOCAA), produced a global mass balance [30, 49], but not a local mass balance.

Eulerian-Lagrangian schemes have been developed to approximate equation (1.1), using Lagrangian characteristic methods for the transport (1.2) and a fixed Eulerian grid for the diffusion (1.3). Included are the Eulerian-Lagrangian localized adjoint methods (ELLAM) [16, 24, 50–52] and the characteristics-mixed method (CMM) [3, 5] and its two-phase variant [31], which are ELLAM schemes but emphasize their development in terms of the local mass constraint. The basic idea is to trace back in time along the characteristics each entire grid element $E$ to $\check{E}$. This creates a tessellation of the domain, and thus all mass can be accounted for locally; that is, all the mass in $\check{E}$ is numerically transported forward into $E$. Eulerian numerical methods based on fixed grids, such as Godunov's method [41], are locally mass conservative by design. They are also automatically volume conserving, since the volumes of the fixed grid elements do not change in time.

For characteristic methods, in the absence of sources, sinks, and external boundaries, the volumes of $E$ and $\check{E}$ agree. However, to trace $\check{E}$ back in time requires tracing each boundary point back, which can be done only in one space dimension (unless perhaps the velocity is particularly simple). So, in practice, one must approximate $\check{E}$ by some simpler shape $\tilde{E}$ by, say, tracing back only the

vertices and midpoints of edges of the element to form a polygon approximation. Almost assuredly the volumes of $\tilde{E}$ and $\check{E}$ will disagree, violating the volume conservation principle (Figure 2.1, left). Although mass may be conserved locally, moving mesh and characteristic methods do not automatically conserve volume, and incorrect local volumes lead to incorrect concentrations, which measure mass per volume. That is, the density is incorrectly approximated and can lead to overshoot or undershoot and seriously degrade the quality of the solution over time.

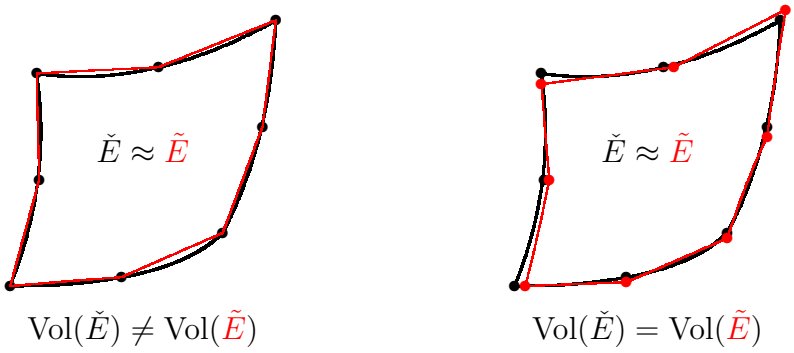


Figure 2.1: Approximation of a trace-back element
Left: The volume imbalance caused by approximating the trace-back $\check{E}$ of element $E$ by a polygon $\tilde{E}$. Right: Correction through small adjustment of the points.

A computationally expensive method was proposed by Chilakapati [20,21] to alleviate the volume discrepancy. He modifies the way in which the velocity $\mathbf{u}$ itself is computed, and then applies CMM to the advection equation.

Arbogast and Huang [4] proposed a much simpler technique that involves postprocessing the approximate trace-back elements $\tilde{E}$ until its volume agrees with $E$ or, equivalently, $\check{E}$, through adjusting the position of the trace-back points (Figure 2.1, right). The method is called the *Volume Corrected Characteristics-Mixed Method* (VCCMM).

A natural approach is to solve a least-squares or other optimization problem to find the minimal trace-back adjustment needed to obtain local volume balance. This approach fails for two reasons: it is prohibitively expensive, and it systematically biases the flow field, since it is not based on the physics of the flow. Great care must be used when adjusting trace-back points to avoid introducing unphysical flow paths into the transport computation. Arbogast and Huang devised algorithms that produce good trace-back regions [4]. The key is to adjust the trace-back points "in time," i.e., along the characteristics [30] (Figure 2.2), where possible. That is, for a time step $J^n$, one traces a point $\mathbf{x}$ at $t^{n+1}$ back to some time $\tau^n \approx t^n$ instead of $t^n$. One proceeds outward from sources of injection by considering an entire layer of elements. One modifies the volume of the layer until it is correct by adjusting simultaneously the points on the "far end" of the layer in time (i.e., along the streamlines). One then adjusts points within the layer lateral to the flow to obtain volume conservation of each individual element. This results in a volume preserving tessellation of the domain.

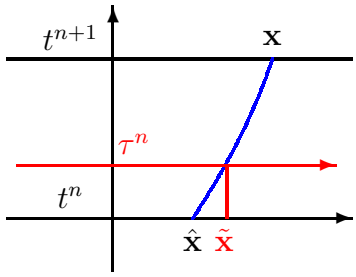


Figure 2.2: Point adjustment along the streamline via time adjustment.

For injection wells, points track backward in time into the well-bore or outside the domain, so one should use a trace-forwarding strategy instead [39]. A brief description of the Volume Correction Algorithm is included in Section 2.2 below.

For small times, the trace-back regions of elements are well defined. However, if the time step is too large, an approximate trace-back region $\tilde{E}$ might

intersect itself. This is unacceptable, and it indicates that the boundary of the true trace-back region $\check{E}$ is convoluted. In this case one should either increase the number of points traced per element edge, or, more practically, reduce the time step. This is in principle the only limitation on the time step.

## 2.2 Local Mass Constraints and the VCCMM Scheme

In the following analysis, we only treat the advective part of the system, i.e., we set $\mathbf{D} = \mathbf{0}$. Furthermore, since $0 < \phi_* \leq \phi(\mathbf{x}) \leq 1$, without losing generality, we assume $\phi(\mathbf{x}) \equiv 1$ for simplicity. That is, we consider variable $\tilde{c} := \phi c$ as the new conserved quantity and introduce the *interstitial* velocity $\mathbf{v} := \mathbf{u}/\phi$, $\tilde{c}_I := \phi c_I$, and $\tilde{q} := q/\phi$. However, we continue to use the notations $c$, $\mathbf{u}$, $c_I$ and $q$. Therefore, the system (2.1)–(2.4) can be reduced to

$$\nabla \cdot (\phi \mathbf{u}) = \phi q \qquad \text{in } \Omega \times J, \tag{2.7}$$

$$c_t + \nabla \cdot (\mathbf{u}c) = q_c := c_I q^+ + c q^- \qquad \text{in } \Omega \times J, \tag{2.8}$$

$$\mathbf{u} \cdot \boldsymbol{\nu} = 0 \qquad \text{on } \partial\Omega \times J, \tag{2.9}$$

$$c(\mathbf{x}, 0) = c^0(\mathbf{x}) \qquad \text{in } \Omega. \tag{2.10}$$

Suppose we have a time interval $J_T := [0, T]$ and a grid $0 = t^0 < t^1 < \cdots < t^N = T$. In one time step $J^n := [t^n, t^{n+1})$, the characteristic trace-back $\check{\mathbf{x}}(t) = \check{\mathbf{x}}(\mathbf{x}, t) = \check{\mathbf{x}}_n(\mathbf{x}, t)$ passing through $(\mathbf{x}, t^{n+1})$ will solve (2.5) and (2.6), unless the particle were to trace to the boundary of the domain, which is excluded by our boundary condition (2.9). (We may omit the subscript of $\check{\mathbf{x}}_n$ if there is no confusion in the context.)

Let $\Omega$ be partitioned into elements $\mathcal{T}_h$ of maximal diameter $h$. Let $E \in \mathcal{T}_h$ be an element of $\Omega$, and define the space-time trace-back region of $E$ as

$$\mathcal{E} = \mathcal{E}_E^n := \{(\check{\mathbf{x}}, t) \in \Omega \times J^n : \check{\mathbf{x}} = \check{\mathbf{x}}_n(\mathbf{x}, t),\ \mathbf{x} \in E\},$$

and the fixed time slice

$$\check{E}(t) = \check{E}_n(t) := \{(\check{\mathbf{x}}, t) \in \Omega \times \{t\} : \check{\mathbf{x}} = \check{\mathbf{x}}_n(\mathbf{x}, t),\ \mathbf{x} \in E\}.$$

Then $E = \check{E}(t^{n+1})$ and the trace-back region of $E$ is $\check{E} = \check{E}(t^n)$.

Let $\boldsymbol{\nu}_{t,\mathbf{x}} := (\nu_t, \boldsymbol{\nu}_{\mathbf{x}})^T$ be the unit outward normal vector to $\partial\mathcal{E}$ and

$$\mathcal{S} = \mathcal{S}_E^n := \{(\check{\mathbf{x}}, t) \in \partial\mathcal{E}_E^n : \check{\mathbf{x}} = \check{\mathbf{x}}_n(\mathbf{x}, t),\ \mathbf{x} \in \partial E\}$$

be the space boundary of the space-time region $\mathcal{E}$. Since $\boldsymbol{\nu}_{t,\mathbf{x}}$ is the unit outward normal vector to $\mathcal{S}$, which is defined by curves tracing in the direction $(1, \mathbf{u})^T$, we have the orthogonality

$$\begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix} \cdot \boldsymbol{\nu}_{t,\mathbf{x}} = 0 \quad \text{on } \mathcal{S}. \tag{2.11}$$

Notice that (2.8) can be rewritten as the space-time divergence

$$\nabla_{t,\mathbf{x}} \cdot \left[ c \begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix} \right] = q_c \quad \text{in } \Omega \times J^n. \tag{2.12}$$

Since $\mathcal{E}$ does not touch $\partial\Omega \times J^n$ by (2.9), applying the divergence theorem and (2.11) to (2.12) gives

$$\iint_{\mathcal{E}} q_c \, d\mathbf{x}\, dt = \iint_{\partial\mathcal{E}} c \begin{pmatrix} 1 \\ \mathbf{u} \end{pmatrix} \cdot \boldsymbol{\nu}_{t,\mathbf{x}} \, d\mathbf{x}\, dt = \int_E c^{n+1} \, d\mathbf{x} - \int_{\check{E}} c^n \, d\mathbf{x},$$

which means the *local mass constraint* is

$$\int_E c^{n+1} \, d\mathbf{x} = \int_{\check{E}} c^n \, d\mathbf{x} + \iint_{\mathcal{E}} q_c \, d\mathbf{x}\, dt, \tag{2.13}$$

where we use superscript $n$ to denote a time dependent function evaluated at time $t^n$.

Due to the approximation of the characteristics (2.5) and approximation of $E$ by a polygon, we actually trace to an approximation $\tilde{E}$ of $\check{E}$. Therefore, the numerical solution

$$c_h^{n+1} \in W_h(\Omega) := \{w \in L^2(\Omega) : w|_E \text{ is a constant for all } E \in \mathcal{T}_h\}$$

is defined on $E$ to be

$$c_{h,E}^{n+1}|E| = \int_{\tilde{E}} c_h^n \, d\mathbf{x} + \iint_{\tilde{\mathcal{E}}} q_{c_h^n} \, d\mathbf{x} \, dt, \tag{2.14}$$

where we define $\tilde{\mathcal{E}}$ in (3.3) later in Chapter 3 as the space-time trace-back region from $E$ to $\tilde{E}$, $|E|$ is the volume of $E$ in the sense of Lebesgue measure in $\mathbb{R}^d$, and $q_{c_h^n} := c_I q^+ + c_h^n q^-$. The numerical solution $c_{h,E}^{n+1}$ is computable since $\tilde{E}$ and $\tilde{\mathcal{E}}$ have replaced $\check{E}$ and $\mathcal{E}$, respectively. We may refer to this method as a *conservative characteristic method*. It is a type of Lagrangian method.

With $c = c_I \equiv \phi$ in (2.13), we have the transport of the single combined fluid (2.7), and

$$|E|_\phi = |\check{E}|_\phi + \iint_{\mathcal{E}} \phi q \, d\mathbf{x} \, dt, \tag{2.15}$$

where $|S|_\phi := \int_S \phi \, d\mathbf{x}$ is the pore volume of a set $S \subset \Omega$. We call (2.15) the *local volume constraint*, since the fluid incompressibly fills the pores. However, it is not likely that

$$|E|_\phi = |\tilde{E}|_\phi + \iint_{\tilde{\mathcal{E}}} \phi q \, d\mathbf{x} \, dt, \tag{2.16}$$

leading to a violation of an important physical principle.

The *volume corrected characteristics-mixed method* [4], which is an Eulerian-Lagrangian method, includes an important procedure for further perturbing the trace-back element $\tilde{E}$ so that (2.16) holds. When $\mathbf{D} = \mathbf{0}$, there is no Eulerian mixed method approximation of the diffusion/dispersion, and so we may refer to the remaining Lagrangian steps as the volume corrected, *fully conservative characteristic method*. We assume that $q = 0$ except in isolated elements of $\mathcal{T}_h$. Assuming for simplicity that the elements are rectangles, given $E \in \mathcal{T}_h$, we trace the four vertices as well as the four midpoints to obtain the unadjusted octagonal polygon $\tilde{E}$. The full algorithm is developed in [4]. A very brief description of the adjustment algorithm follows.

**The Volume Correction Algorithm**

**Point adjustment in time.** A trace-back point may be adjusted *in time* by a small amount [30], along the characteristics in the direction of the flow field. As we will see, the effect is to convert spatial errors into time errors. Moreover, in this way, no bias is introduced into the *direction* of the flow. This time adjustment is needed in Steps 1 and 2 below.

**Step 1: Forward trace out of injection wells.** Trace forward (not backward, see, e.g., (3.11)–(3.12)) the injection wells [39], and then adjust the trace-forward boundary in time according to the well volume constraint.

**Step 2: Ring adjustment.** Between the wells, starting adjacent to the injection well and moving towards the production wells, entire rings of elements are adjusted in time so as to have the correct volume. Assuming the trace-back ring edge closest to the injector has been adjusted, the points on the far edge are adjusted simultaneously.

**Step 3: Individual element adjustment.** Within an adjusted ring of elements, individual elements are adjusted to have the correct volume by traversing the ring, starting from a no-flow boundary if one intersects the ring. This is accomplished by a transverse movement of the midpoints (not a time adjustment).

For consistency of the trace-back tessellation, we tacitly assume that the time step is restricted so that the trace-back elements $\tilde{E}$ do not self intersect. Moreover, we assume that no sink traces all the way to a source within a single time step.

## 2.3 Stability Analysis

In this section, for simplicity, we assume $\phi(\mathbf{x}) \equiv 1$ in $\Omega$ as in (2.8)–(2.10). Let vector $\mathbf{c}_h^n := (c_{h,E}^n)_{E \in \mathcal{T}_h} \in \mathbb{R}^{N_h}$ for $0 \leq n \leq N$, where $N_h$ is the number of

elements in the mesh $\mathcal{T}_h$. Assume we have isolated injection wells and production wells in the domain $\Omega$. Let $\mathcal{T}_{h,P} \subset \mathcal{T}_h$ be the collection of elements which represent locations of production wells. We derive the scheme of VCCMM in a vector form for $\mathbf{c}_h^n$ below.

### 2.3.1 Derivation of VCCMM in a vector form

Since we use a piecewise constant function $c_h^n \in W_h(\Omega)$ to approximate the solution, the scheme of the VCCMM (2.14) in an integral form is reduced to

$$\begin{aligned} c_{h,E}^{n+1}|E| &= \sum_{F\in\mathcal{T}_h} c_{h,F}^n |\tilde{E}^n \cap F| + \sum_{F\in\mathcal{T}_h} c_{h,F}^n \iint_{\tilde{\mathcal{E}}_E^n \bigcap I_F^n} q^-\, d\mathbf{x}\, dt + \iint_{\tilde{\mathcal{E}}_E^n} c_I q^+\, d\mathbf{x}\, dt \\ &= \sum_{F\in\mathcal{T}_h} c_{h,F}^n \left( |\tilde{E}^n \cap F| + \iint_{\tilde{\mathcal{E}}_E^n \bigcap I_F^n} q^-\, d\mathbf{x}\, dt \right) + \iint_{\tilde{\mathcal{E}}_E^n} c_I q^+\, d\mathbf{x}\, dt, \end{aligned} \tag{2.17}$$

where $I_F^n$ is the space-time cylinder $F \times J^n$. When $E \notin \mathcal{T}_{h,P}$, $q \geq 0$ in $\tilde{\mathcal{E}}_E^n$, so (2.17) is reduced to

$$c_{h,E}^{n+1}|E| = \sum_{F\in\mathcal{T}_h} c_{h,F}^n |\tilde{E}^n \cap F| + \iint_{\tilde{\mathcal{E}}_E^n} c_I q^+\, d\mathbf{x}\, dt, \quad E \notin \mathcal{T}_{h,P}. \tag{2.18}$$

When $E \in \mathcal{T}_{h,P}$, $E \subset \tilde{E}^n$ and $I_E^n \subset \tilde{\mathcal{E}}_E^n$, since the trace-back of a production well boundary expands. Also, we have $q = 0$ in $\tilde{\mathcal{E}}_E^n \cap I_F^n$ if $\tilde{\mathcal{E}}_E^n \cap I_F^n \neq \emptyset$ and $F \neq E$, since we assume that no sink traces all the way to a well within a single time step. Then (2.17) is reduced to

$$\begin{aligned} c_{h,E}^{n+1}|E| = \sum_{F\neq E} c_{h,F}^n |\tilde{E}^n \cap F| + c_{h,E}^n \left( |E| + \iint_{I_E^n} q^-\, d\mathbf{x}\, dt \right) \\ + \iint_{\tilde{\mathcal{E}}_E^n} c_I q^+\, d\mathbf{x}\, dt, \quad E \in \mathcal{T}_{h,P}. \end{aligned} \tag{2.19}$$

Notice that the coefficient of $c_{h,E}^n$ in (2.19) is

$$V_E^n := |E| + \iint_{I_E^n} q^-\, d\mathbf{x}\, dt, \tag{2.20}$$

which is the remaining volume of the combined fluids in the production well in $E$ at time $t^n$, so $V_E^n$ should be non-negative in physical terms, although $V_E^n$ could be negative numerically if we have a strong production rate in $E$. So when $V_E^n < 0$, we modify (2.19) by (1) setting the remaining volume of the combined fluids in $E$ to be $(V_E^n)^+ = 0$ and (2) reducing a certain volume of fluids from each nearby element by a proportion such that the local volume constraint (2.16) still holds. That is, for each $E \in \mathcal{T}_{h,P}$, we consider the identity

$$V_E^n = (V_E^n)^+ + \sum_{F \neq E} \frac{|\tilde{E}^n \cap F|}{|\tilde{E}^n \setminus E|} (V_E^n)^-$$

and modify (2.19) to be

$$\begin{aligned} c_{h,E}^{n+1}|E| = \sum_{F \neq E} c_{h,F}^n \left( |\tilde{E}^n \cap F| + \frac{|\tilde{E}^n \cap F|}{|\tilde{E}^n \setminus E|} (V_E^n)^- \right) + c_{E,h}^n (V_E^n)^+ \\ + \iint_{\tilde{\mathcal{E}}_E^n} c_I q^+ \, d\mathbf{x} \, dt, \quad E \in \mathcal{T}_{h,P}. \end{aligned} \tag{2.21}$$

Combining (2.18) and (2.21) gives the scheme of VCCMM in a vector form

$$\mathbf{c}_h^{n+1} = \mathbf{A}_h^n \mathbf{c}_h^n + \mathbf{b}_h^n, \tag{2.22}$$

where matrix $\mathbf{A}_h^n = (A_{h,E,F}^n) \in \mathbb{R}^{N_h \times N_h}$ and vector $\mathbf{b}_h^n = (b_{h,E}^n) \in \mathbb{R}^{N_h}$ are given by

$$A_{h,E,F}^n := \begin{cases} \dfrac{|\tilde{E}^n \cap F|}{|E|}, & E \notin \mathcal{T}_{h,P}, \\ \dfrac{|\tilde{E}^n \cap F|}{|E|} \left( 1 + \dfrac{(V_E^n)^-}{|\tilde{E}^n \setminus E|} \right), & E \in \mathcal{T}_{h,P},\ F \neq E, \\ \dfrac{(V_E^n)^+}{|E|}, & E = F \in \mathcal{T}_{h,P}, \end{cases} \tag{2.23}$$

$$b_{h,E}^n := \frac{1}{|E|} \iint_{\tilde{\mathcal{E}}_E^n} c_I q^+ \, d\mathbf{x} \, dt. \tag{2.24}$$

### 2.3.2 Stability of VCCMM

**Lemma 2.3.1.** *The matrix* $\mathbf{A}_h^n$ *defined in* (2.23) *as a vector operator does not increase the* $l^\infty$*-norm of a vector, i.e., for any* $\mathbf{c} \in \mathbb{R}^{N_h}$,

$$|\mathbf{A}_h^n \mathbf{c}|_\infty \leq |\mathbf{c}|_\infty.$$

*Proof.* First we will show each entry $A_{h,E,F}^n \geq 0$. By (2.23), we only need to show $|\tilde{E}^n \setminus E| + (V_E^n)^- \geq 0$ for $E \in \mathcal{T}_{h,P}$ and $F \neq E$. Actually,

$$\begin{aligned} |\tilde{E}^n \setminus E| + (V_E^n)^- &\geq |\tilde{E}^n \setminus E| + \iint_{I_E^n} q^- \, d\mathbf{x}\, dt \\ &= |\tilde{E}^n| - |E| + \iint_{\tilde{\mathcal{E}}_E^n} q \, d\mathbf{x}\, dt = 0, \end{aligned}$$

where we obtain the last equality by the local volume constraint (2.16).

Then we will show each row sum of $\mathbf{A}_h^n$

$$\sum_{F \in \mathcal{T}_h} A_{h,E,F}^n \leq 1. \tag{2.25}$$

By (2.23) and (2.16), we compute when $E \notin \mathcal{T}_{h,P}$,

$$\sum_{F \in \mathcal{T}_h} A_{h,E,F}^n = \frac{|\tilde{E}^n|}{|E|} = \frac{1}{|E|} \left( |E| - \iint_{\tilde{\mathcal{E}}_E^n} q \, d\mathbf{x}\, dt \right) \leq 1.$$

and when $E \in \mathcal{T}_{h,P}$,

$$\begin{aligned} \sum_{F \in \mathcal{T}_h} A_{h,E,F}^n &= \frac{|\tilde{E}^n \setminus E|}{|E|} \left( 1 + \frac{(V_E^n)^-}{|\tilde{E}^n \setminus E|} \right) + \frac{(V_E^n)^+}{|E|} \\ &= \frac{1}{|E|} (|\tilde{E}^n \setminus E| + V_E^n) = \frac{1}{|E|} \left( |\tilde{E}^n| + \iint_{I_E^n} q^- \, d\mathbf{x}\, dt \right) \\ &= \frac{1}{|E|} \left( |\tilde{E}^n| + \iint_{\tilde{\mathcal{E}}_E^n} q \, d\mathbf{x}\, dt \right) = 1, \end{aligned}$$

so we obtain (2.25).

For any $\mathbf{c} \in \mathbb{R}^{N_h}$ and $E \in \mathcal{T}_h$, we have

$$\left|(\mathbf{A}_h^n \mathbf{c})_E\right| = \left|\sum_{F \in \mathcal{T}_h} A_{h,E,F}^n c_F\right| \leq \sum_{F \in \mathcal{T}_h} A_{h,E,F}^n |c_F| \leq \sum_{F \in \mathcal{T}_h} A_{h,E,F}^n |\mathbf{c}|_\infty \leq |\mathbf{c}|_\infty,$$

which means

$$|\mathbf{A}_h^n \mathbf{c}|_\infty = \max_{E \in \mathcal{T}_h} \left|(\mathbf{A}_h^n \mathbf{c})_E\right| \leq |\mathbf{c}|_\infty.$$

□

**Theorem 2.3.2** (Stability of VCCMM)**.** *The scheme of VCCMM given by* (2.22) *is stable. That is, if* $\mathbf{c}_h^n$ $(0 \leq n \leq N)$ *satisfies scheme* (2.22) *in time* $J_T$ *with an initial approximation* $\mathbf{c}_h^0$*, and* $\tilde{\mathbf{c}}_h^n$ $(0 \leq n \leq N)$ *satisfies the perturbed scheme*

$$\tilde{\mathbf{c}}_h^{n+1} = \mathbf{A}_h^n \tilde{\mathbf{c}}_h^n + \mathbf{b}_h^n + \Delta t^n \boldsymbol{\delta}_h^n \tag{2.26}$$

*in time* $J_T$ *with an initial approximation* $\tilde{\mathbf{c}}_h^0$*, where* $\boldsymbol{\delta}_h^n \in \mathbb{R}^{N_h}$ *is a perturbation at time step* $J^n$*, then the following error estimate holds:*

$$\max_{0 \leq n \leq N} |\tilde{\mathbf{c}}_h^n - \mathbf{c}_h^n|_\infty \leq |\tilde{\mathbf{c}}_h^0 - \mathbf{c}_h^0|_\infty + T \max_{0 \leq n \leq N} |\boldsymbol{\delta}_h^n|_\infty. \tag{2.27}$$

*Proof.* Subtracting (2.22) from (2.26), we have

$$\tilde{\mathbf{c}}_h^{n+1} - \mathbf{c}_h^{n+1} = \mathbf{A}_h^n (\tilde{\mathbf{c}}_h^n - \mathbf{c}_h^n) + \Delta t^n \boldsymbol{\delta}_h^n. \tag{2.28}$$

Taking $l^\infty$-norm on both sides of (2.28) and using Lemma 2.3.1 give

$$\begin{aligned} |\tilde{\mathbf{c}}_h^{n+1} - \mathbf{c}_h^{n+1}|_\infty &\leq |\mathbf{A}_h^n (\tilde{\mathbf{c}}_h^n - \mathbf{c}_h^n)|_\infty + \Delta t^n |\boldsymbol{\delta}_h^n|_\infty \\ &\leq |\tilde{\mathbf{c}}_h^n - \mathbf{c}_h^n|_\infty + \Delta t^n \max_{0 \leq n \leq N} |\boldsymbol{\delta}_h^n|_\infty. \end{aligned} \tag{2.29}$$

Iterating (2.29) for $n$ gives

$$|\tilde{\mathbf{c}}_h^n - \mathbf{c}_h^n|_\infty \leq |\tilde{\mathbf{c}}_h^0 - \mathbf{c}_h^0|_\infty + t^n \max_{0 \leq n \leq N} |\boldsymbol{\delta}_h^n|_\infty$$

and we obtain (2.27). □

# Chapter 3

# Convergence Analysis of VCCMM

Without volume adjustment, in [5], it is proven that the CMM is first order convergent in the mesh spacing parameter $h$ with a non-degenerate diffusion-dispersion tensor. Without diffusion-dispersion (i.e., $\mathbf{D} = \mathbf{0}$), due to projection error accumulation [42], piecewise discontinuous constant approximations can be only $\mathcal{O}(h/\sqrt{\Delta t}+h+(\Delta t)^r)$, where $r$ is related to the accuracy of the characteristic tracing (see Remark 3.5.2 below).

We use a key idea introduced by Arbogast and Wheeler [5], wherein it was noted that an analysis of inexact characteristic tracing, i.e., approximation of the solution to (2.5)–(2.6), could be made if one views the approximate tracing as arising from exact tracing through a perturbed velocity field. In addition, for the volume correction of VCCMM, we will construct this perturbed velocity $\tilde{\mathbf{u}}$ such that each trace-back of element $E \in \mathcal{T}_h$ is the volume corrected $\tilde{E}$. That is, we replace $\mathbf{u}$ in (2.5)–(2.6) by $\tilde{\mathbf{u}}$ and solve for $\tilde{\mathbf{x}}(t) = \tilde{\mathbf{x}}(\mathbf{x}, t) = \tilde{\mathbf{x}}_n(\mathbf{x}, t)$ the approximate tracing

$$\tilde{\mathbf{x}}_t = \tilde{\mathbf{u}}(\tilde{\mathbf{x}}, t), \quad t \in J^n, \tag{3.1}$$

$$\tilde{\mathbf{x}}(t^{n+1}) = \mathbf{x}. \tag{3.2}$$

Then, for each $E \in \mathcal{T}_h$, we can define the numerical space-time region

$$\tilde{\mathcal{E}} = \tilde{\mathcal{E}}_E^n = \{(\tilde{\mathbf{x}}, t) \in \Omega \times J^n : \tilde{\mathbf{x}} = \tilde{\mathbf{x}}_n(\mathbf{x}, t),\ \mathbf{x} \in E\}, \tag{3.3}$$

and the numerical fixed time slice

$$\tilde{E}(t) = \tilde{E}_n(t) = \{(\tilde{\mathbf{x}}, t) \in \Omega \times \{t\} : \tilde{\mathbf{x}} = \tilde{\mathbf{x}}_n(\mathbf{x}, t),\ \mathbf{x} \in E\},$$

for which $E = \tilde{E}(t^{n+1})$ and volume corrected trace-back region of $E$ is $\tilde{E} = \tilde{E}(t^n)$.

However, the existence of $\tilde{\mathbf{u}}$ and estimate of the error $(\mathbf{u} - \tilde{\mathbf{u}})$ present the major difficulty. The construction of $\tilde{\mathbf{u}}$ will be given in Section 3.5. For now, we simply make the following assumption. We use $\|\cdot\|_{p,S}$ to denote the norm of $L^p(S)$ and we may omit $S$ if $S = \Omega$ or $\Omega \times J_T$.

*Assumption* 3.0.1 (Perturbed velocity field). The velocity field $\mathbf{u} = \mathbf{u}(\mathbf{x}, t) \in C^1(\Omega \times J_T)$ has divergence $\nabla \cdot \mathbf{u}(\cdot, t)$ uniformly Lipschitz continuous in time $J_T$, i.e.,

$$|\nabla \cdot \mathbf{u}(\mathbf{x}, t) - \nabla \cdot \mathbf{u}(\mathbf{y}, t)| \le L|\mathbf{x} - \mathbf{y}| \quad \text{for all } \mathbf{x}, \mathbf{y} \in \Omega,\ t \in J_T, \tag{3.4}$$

where $L > 0$ is a constant independent of $\mathbf{x}$, $\mathbf{y}$, and $t$. There exists a locally conservative velocity field $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\mathbf{x}, t)$ on $\Omega \times J_T$ such that

$$\tilde{\mathbf{u}} \cdot \boldsymbol{\nu} = 0 \text{ on } \partial\Omega \times J_T, \tag{3.5}$$

each trace-back polygon $\tilde{E}$ satisfies the local volume constraint (2.16), and

$$\|\mathbf{u} - \tilde{\mathbf{u}}\|_\infty + \|\nabla \cdot \mathbf{u} - \nabla \cdot \tilde{\mathbf{u}}\|_\infty \le C(h + (\Delta t)^r), \tag{3.6}$$

where $C$ and $r > 0$ are constants independent of $h$ and $\Delta t$.

Assume $c_h^0$ is a given initial approximation of $c^0$. In each time step $J^n$, now we consider $c_h$ is a solution to the perturbed system

$$(c_h)_t + \nabla \cdot (c_h \tilde{\mathbf{u}}) = q_{c_h} \quad \text{in } \Omega \times J^n, \tag{3.7}$$

$$c_h(\mathbf{x}, t^n) = c^n(\mathbf{x}) \qquad \text{in } \Omega, \tag{3.8}$$

and we define the update at $t^{n+1}$ as

$$c_h^{n+1}(\mathbf{x}) := P_h c_h(\mathbf{x}, t^{n+1}-) = P_h c_h^{n+1-}(\mathbf{x}), \tag{3.9}$$

where the $L^2$-projection operator $P_h$ is defined as

$$(P_h f, w) = (f, w) \text{ for all } w \in W_h(\Omega). \tag{3.10}$$

The rest of the chapter is organized as follows. Section 3.1 gives an analytical representation of the weak solution and introduces the entropy inequality. Section 3.2 lists and proves some properties of the weak solution and the numerical solution that are relevant to our purposes. Section 3.3 introduces an approximation of $L^1$-errors and proves some properties that play an important role in the proof of convergence. Section 3.4 gives the convergence result for the method. Section 3.5 gives the existence and an error estimate of the perturbed velocity field, which presents the major difficulty of our overall proof. Summary and concluding remarks are given in the last section.

## 3.1 An Analytical Representation of the Weak Solution and the Entropy Inequality

Taking advantage of the linear structure of transport equation (2.8), as is well known, we can actually solve system (2.8)–(2.10) analytically by integration along characteristics. Let $\hat{\mathbf{x}} = \hat{\mathbf{x}}(\mathbf{x}, t)$ be the trace-forward characteristics of $\mathbf{u}$, i.e.,

$$\hat{\mathbf{x}}_t = \hat{\mathbf{u}} \qquad \text{in } \Omega \times J_T, \tag{3.11}$$

$$\hat{\mathbf{x}}(\mathbf{x}, 0) = \mathbf{x} \quad \text{in } \Omega, \tag{3.12}$$

where $\hat{f}(\mathbf{x}, t) := f(\hat{\mathbf{x}}(\mathbf{x}, t), t)$ is the evaluation along trace-forward characteristics for a generic scalar or vector valued function $f$.

**Lemma 3.1.1** (Analytical representation)**.** *Let* $\mathbf{u}$ *be a smooth velocity field on the domain* $\Omega \times J_T$ *and* $\hat{\mathbf{x}}$ *be the trace-forward characteristics of* $\mathbf{u}$ *defined in* (3.11)–(3.12)*. For any* $t \in J_T$*, assume* $\hat{\mathbf{x}}(\cdot, t)$ *is a diffeomorphism in* $\Omega$*, and denote the inverse as* $\check{\mathbf{x}}(\cdot, t)$*. Then the weak solution to system* (2.8)–(2.10) *evaluated along characteristics is given by*

$$\hat{c} = F_0 + F_1 c^0, \tag{3.13}$$

*where*

$$F_1(\mathbf{x},t) := \exp\left(\int_0^t (q^- - \nabla\cdot\mathbf{u})^\wedge(\mathbf{x},s)\,ds\right), \tag{3.14}$$

$$F_0(\mathbf{x},t) := \int_0^t f_0(\mathbf{x},t,s)\,ds, \tag{3.15}$$

$$f_0(\mathbf{x},t,s) := (c_I q^+)^\wedge(\mathbf{x},s)\frac{F_1(\mathbf{x},t)}{F_1(\mathbf{x},s)}. \tag{3.16}$$

*Proof.* Rearrange (2.8), and we have

$$c_t + \mathbf{u}\cdot\nabla c = c_I q^+ + (q^- - \nabla\cdot\mathbf{u})c.$$

Notice that

$$(\hat{c})_t = (c(\hat{\mathbf{x}}(\mathbf{x},t),t))_t = \widehat{c_t} + \hat{\mathbf{u}}\cdot\nabla\hat{c},$$

so $\hat{c}$ solves the well-posed initial value problem of an ordinary differential equation

$$\begin{aligned} &(\hat{c})_t = (c_I q^+)^\wedge + (q^- - \nabla\cdot\mathbf{u})^\wedge\hat{c} && \text{in } \Omega\times J_T,\\ &\hat{c}(\mathbf{x},0) = c^0(\mathbf{x}) && \text{in } \Omega. \end{aligned}$$

Then we obtain (3.13) by solving the ordinary differential equation above. □

The analytical representation implies the existence and uniqueness of the weak solution.

**Corollary 3.1.2** (Existence and uniqueness)**.** *If the trace-forward characteristics* $\hat{\mathbf{x}}$ *of* $\mathbf{u}$ *form a diffeomorphism, then there exists a unique weak solution* $c$ *to system* (2.7)–(2.10) *given by* (3.13)*.*

By the theory of conservation laws, the weak solution $c = c(\mathbf{x},t)$ also satisfies a stability condition, which is called the *entropy inequality* or *entropy admissibility condition*, relative to a convex entropy $\eta$; that is,

$$\eta_t + \nabla\cdot\mathbf{Q} \le H$$

in the sense of distributions, i.e.,

$$(\eta^n, \varphi^n) - (\eta^{n+1-}, \varphi^{n+1}) + \int_{J^n} (\eta, \varphi_t)\, dt \tag{3.17}$$
$$- \int_{J^n} \langle \mathbf{Q} \cdot \boldsymbol{\nu}, \varphi \rangle_{\partial\Omega}\, dt + \int_{J^n} (\mathbf{Q}, \nabla\varphi)\, dt + \int_{J^n} (H, \varphi)\, dt \geq 0$$

for any non-negative test function $\varphi = \varphi(\mathbf{x}, t) \in C^\infty(\Omega \times J^n)$. Any convex function $\eta = \eta(c)$ may serve as an entropy [23, pp. 54], with the associated entropy flux $\mathbf{Q}$ and entropy production $H$ computed by

$$\mathbf{Q} = \eta\, \mathbf{u} \quad \text{and} \quad H = \eta' q_c + (\eta - \eta' c) \nabla \cdot \mathbf{u}.$$

Note that the term involving $\mathbf{Q} \cdot \boldsymbol{\nu}$ in (3.17) vanishes by the boundary condition (2.9). In general, the entropy solution is the weak solution which is physically relevant. In our case, there is only one solution, and we will use (3.17) freely.

## 3.2 Properties of the Weak Solution

It is well known from the theory of scalar conservation laws, with a flux $\mathbf{F}$ in the canonical form

$$c_t + \nabla \cdot \mathbf{F}(c) = 0 \quad \text{in } \mathbb{R}^d \times \mathbb{R}^+, \tag{3.18}$$
$$c(\mathbf{x}, 0) = c^0(\mathbf{x}) \qquad \text{in } \mathbb{R}^d, \tag{3.19}$$

that the law has reached a state of virtual completeness, such as $L^1$-contraction, uniqueness, $L^\infty$-monotonicity, uniform boundedness, and total variation diminishing (TVD) properties of the entropy solution [23, pp. 126–142].

It should be noted that our transport equation (2.8) is similar to, but not a subcase of, the canonical form (3.18), which is homogenous and the flux $\mathbf{F}$ does not explicitly depend on spatial and time variables, but only on the conserved quantity $c$. In this section, we prove some properties of the weak solution to the system (2.7)–(2.10) that are relevant to our purposes in the following analysis.

### 3.2.1 Uniform boundedness

Physically, the tracer mass comes from the initial state and the injected concentration as time proceeds. Indeed, by the analytical representation (3.13), it is easy to see the uniform boundedness of the weak solution.

**Lemma 3.2.1** (Boundedness of the weak solution)**.** *If* $c^0 \in L^\infty(\Omega)$ *and* $c_I, q \in L^\infty(\Omega \times J_T)$*, the weak solution* $c$ *to the system* (2.7)–(2.10) *is uniformly* $L^\infty$*-bounded and* $L^1$*-bounded in* $\Omega \times J_T$*.*

*Proof.* By the analytical representation (3.13), it is easy to see the uniform $L^\infty$-boundedness of $c$. Then the $L^1$-boundedness of $c$ follows due to the boundedness of the space-time domain $\Omega \times J_T$. □

**Lemma 3.2.2** (Boundedness of the numerical solution)**.** *If* $c_h^0 \in L^\infty(\Omega)$ *and* $c_I$*,* $q \in L^\infty(\Omega \times J_T)$*, the numerical solution* $c_h$ *to the system* (3.7)–(3.9) *is uniformly* $L^\infty$*-bounded and* $L^1$*-bounded in* $\Omega \times J_T$*.*

*Proof.* Notice that the $L^2$-projection operator defined in (3.10) increases neither the $L^\infty$- nor $L^1$-norm of a function, so we can perform a similar argument as in Lemma 3.2.1 for $c_h$ defined in (3.7)–(3.9) in each time step $J^n$ to complete the proof. □

### 3.2.2 Boundedness of the total variation (TVB)

Variations of solutions play an important role in hyperbolic differential equations. In this subsection, we list and prove some basic properties of functions of bounded variation, prove the total variation boundedness (TVB) property of the weak solution, and make an assumption on the $L^1$-TVB property of the numerical solution.

#### 3.2.2.1 Properties of functions of bounded variation

The total variation of a function $f$ on $\Omega$ is defined by

$$|f|_{BV(\Omega)} = \sup_{\boldsymbol{\varphi}} (f, \nabla \cdot \boldsymbol{\varphi})_{L^2(\Omega)}, \tag{3.20}$$

where the supremum is taken for all vector-valued functions $\boldsymbol{\varphi} = (\varphi_1, \cdots, \varphi_d)^T \in [C_c^\infty(\Omega)]^d$ with $\|\boldsymbol{\varphi}\|_\infty := \max_{1\le i\le d} \|\varphi_i\|_\infty \le 1$. We denote $BV(\Omega) := \{f \in L^1(\Omega) : |f|_{BV(\Omega)} < \infty\}$ to be the set of $L^1$ functions of bounded variation on $\Omega$. Then $|\cdot|_{BV(S)}$ is a semi-norm on $BV(S)$, and we may omit $S$ if $S = \Omega$. If $f \in W^{1,1}(\Omega)$, integrating by parts, we have

$$|f|_{BV} = \|\nabla f\|_1 := \sum_{i=1}^{d} \|\partial_i f\|_1 < \infty, \tag{3.21}$$

and so $W^{1,1}(\Omega) \subset BV(\Omega) \subset L^1(\Omega)$.

**Proposition 3.2.3.** *If the domain $\Omega$ has a partition $\mathcal{T}$, then for any $f \in BV(\Omega)$,*

$$\sum_{E\in\mathcal{T}} |f|_{BV(E)} \le |f|_{BV(\Omega)}.$$

Proposition 3.2.3 is trivial to prove by definition (3.20).

**Proposition 3.2.4** (Lower semicontinuity)**.** *The $BV$ seminorm is lower semicontinuous with respect to the $L^1$-topology, i.e., if $f_j \to f$ in $L^1(\Omega)$, then*

$$|f|_{BV} \le \liminf_{j\to\infty} |f_j|_{BV}.$$

*Proof.* See [37, pp. 7]. □

**Proposition 3.2.5** (Approximation by smooth functions)**.** *For any $f \in BV(\Omega)$, there exists a sequence $\{f_j\}$ in $C^\infty(\Omega)$ such that $f_j \to f$ in $L^1(\Omega)$ and $|f_j|_{BV} \to |f|_{BV}$.*

*Proof.* See [37, pp. 14]. □

**Proposition 3.2.6** (Product rule)**.** *For any $f \in BV(\Omega)$ and $g \in W^{1,\infty}(\Omega)$, the product $fg \in BV(\Omega)$, and*

$$|fg|_{BV} \leq |f|_{BV}\|g\|_\infty + \|f\|_1\|\nabla g\|_\infty. \tag{3.22}$$

*Proof.* First suppose $f \in C^\infty(\Omega)$. By taking $L^1$-norms on both sides of the identity

$$\nabla(fg) = g\nabla f + f\nabla g,$$

we obtain (3.22). Now for general $f \in BV(\Omega)$, by Proposition 3.2.5, there is a sequence $\{f_j\}$ in $C^\infty(\Omega)$ such that $f_j \to f$ in $L^1(\Omega)$ and $|f_j|_{BV} \to |f|_{BV}$. Now $f_j g \to fg$ in $L^1(\Omega)$. By Proposition 3.2.4 and (3.22) for smooth functions, we have

$$\begin{aligned} |fg|_{BV} &\leq \liminf_{j\to\infty} |f_j g|_{BV} \leq \liminf_{j\to\infty} (|f_j|_{BV}\|g\|_\infty + \|f_j\|_1\|\nabla g\|_\infty) \\ &= |f|_{BV}\|g\|_\infty + \|f\|_1\|\nabla g\|_\infty. \end{aligned}$$

□

**Proposition 3.2.7** (Composition rule)**.** *For any $f \in BV(\Omega)$ and diffeomorphism $g$ on $\Omega$, the composition $f \circ g \in BV(\Omega)$, and*

$$|f \circ g|_{BV} \leq \|\nabla g\|_\infty \|\det(\nabla g^{-1})\|_\infty |f|_{BV}. \tag{3.23}$$

*Proof.* For $f \in C^\infty(\Omega)$, by taking $L^1$-norms on both sides of the identity

$$\nabla(f \circ g) = \nabla g(\nabla f) \circ g$$

and changing variables, we obtain (3.23). The result for general $f \in BV(\Omega)$ follows from Propositions 3.2.4 and 3.2.5 as in the previous proof. □

**Proposition 3.2.8** (Difference quotient)**.** *If the domain $\Omega$ is convex, then the integral of the difference quotient is bounded by the total variation. That is, for any $f \in BV(\Omega)$,*

$$\sup_{\mathbf{y}\neq 0} \|D_{\mathbf{y}} f\|_{1,\Omega_{\mathbf{y}}} \leq |f|_{BV(\Omega)}, \tag{3.24}$$

*where* $D_{\mathbf{y}} := |\mathbf{y}|^{-1}(T_{\mathbf{y}} - I)$ *is the difference quotient operator with the translation operator* $T_{\mathbf{y}}$ *defined by*

$$(T_{\mathbf{y}} f)(\mathbf{x}) = f(\mathbf{x} + \mathbf{y}),$$

*and* $\Omega_{\mathbf{y}} = \Omega \cap (\Omega - \{\mathbf{y}\})$ *is the restricted domain on which the integral is well defined.*

*Proof.* By Proposition 3.2.5, we only need to show (3.24) for $f \in C^\infty(\Omega)$. For any $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y} \neq 0$, and $\mathbf{x} \in \Omega_{\mathbf{y}}$, if $\Omega_{\mathbf{y}}$ is not empty, we have the identity

$$f(\mathbf{x} + \mathbf{y}) - f(\mathbf{x}) = \int_0^1 \nabla f(\mathbf{x} + s\mathbf{y}) \cdot \mathbf{y}\, ds.$$

Taking norms on both sides and integrating with respect to $\mathbf{x} \in \Omega_{\mathbf{y}}$, we have

$$\begin{aligned} \|D_{\mathbf{y}} f\|_{1,\Omega_{\mathbf{y}}} &= \frac{1}{|\mathbf{y}|} \int_{\Omega_{\mathbf{y}}} \left| \int_0^1 \nabla f(\mathbf{x} + s\mathbf{y}) \cdot \mathbf{y}\, ds \right| d\mathbf{x} \\ &\leq \int_0^1 \int_{\Omega_{\mathbf{y}}} |\nabla f(\mathbf{x} + s\mathbf{y})|\, d\mathbf{x}\, ds \leq \|\nabla f\|_{1,\Omega} = |f|_{BV(\Omega)}. \end{aligned}$$

□

#### 3.2.2.2 TVB property

Since we have a balance law, i.e., a conservation law in an inhomogeneous form (2.8), unfortunately, we cannot expect it obeys the total variation diminishing (TVD) property in general. However, since we study the solution in a bounded time interval $J_T$ and the transport equation (2.8) is linear, the physical behavior of the solution should continuously change as time proceeds. It is natural to expect the solution is TVB in $J_T$ under some regularity assumptions of the data in the system (2.7)–(2.10). Denote

$$\begin{aligned} V(\Omega) &:= L^\infty(\Omega) \cap W^{1,1}(\Omega), \\ V(J_T^k; \Omega) &:= L^\infty(\Omega \times J_T^k) \cap C(J_T^k; W^{1,1}(\Omega)), \end{aligned}$$

where $k$ is an positive integer. Note $f \in C(J_T^k; W^{m,p}(\Omega))$ means $f(\cdot, t_1, \cdots, t_k) \in W^{m,p}(\Omega)$ and $\|f(\cdot, t_1, \cdots, t_k)\|_{W^{m,p}(\Omega)}$ is continuous for each $t_j \in J_T$.

*Assumption* 3.2.1 (Regularity of data). Velocity field $\mathbf{u} \in C^1(\Omega \times J_T)$ with diffeomorphic characteristics $\hat{\mathbf{x}}$, $\nabla \cdot \mathbf{u}$ satisfies the uniform Lipschitz condition (3.4), $q \in C(J_T; W^{1,\infty}(\Omega))$, $c^0 \in V(\Omega)$, and $c_I \in V(J_T; \Omega)$.

*Assumption* 3.2.2 (Regularity of initial approximation). $c_h^0 \in L^\infty(\Omega) \cap BV(\Omega)$.

Furthermore, we impose the following assumptions on the time and space domain discretizations.

*Assumption* 3.2.3 (Regularity of time discretization). The time grid $0 = t^0 < t^1 < \cdots < t^N = T$ of $J_T$ is regular, i.e., there exists a constant $\lambda_1 > 0$ such that

$$\Delta t \leq \lambda_1 \inf_n \Delta t^n,$$

where $\Delta t^n := t^{n+1} - t^n$ and $\Delta t := \sup_n \Delta t^n$.

*Assumption* 3.2.4 (Shape regularity of domain discretization). The mesh $\mathcal{T}_h$ of bounded domain $\Omega$ is convex and regular, i.e., each element $E \in \mathcal{T}_h$ is convex, and there exists a constant $\lambda_2 > 0$ such that

$$\sup_{E \in \mathcal{T}_h} \frac{h_E}{\rho_E} \leq \lambda_2,$$

where $h_E$ and $\rho_E$ are the outer and inner diameters of element $E$, respectively, and the mesh spacing parameter $h := \sup_{E \in \mathcal{T}_h} h_E < \infty$.

**Lemma 3.2.9** (TVB of the weak solution)**.** *Let Assumption* 3.2.1 *hold. Then the weak solution $c$ to the system* (2.8)–(2.10) *is TVB to time $T$. Moreover,*

$$|c(\cdot, t)|_{BV} \leq C_0 t + e^{C_1 t} |c^0|_{BV}, \tag{3.25}$$

*where $C_0 > 0$ and $C_1 > 0$ are constants independent of $t$.*

*Proof.* By Assumption 3.2.1, we see from (3.14)–(3.16), that $F_1 \in W^{1,\infty}(\Omega)$, $f_0 \in V(J_T^2; \Omega)$, and $F_0 \in V(J_T; \Omega)$. By the analytical representation (3.13), we have

$$|\hat{c}(\cdot, t)|_{BV} \leq |F_0(\cdot, t)|_{BV} + |(F_1 c^0)(\cdot, t)|_{BV}, \tag{3.26}$$

and

$$|F_0(\cdot,t)|_{BV} = \|\nabla F_0(\cdot,t)\|_1 = \left\| \int_0^t \nabla f_0(\cdot,t,s)\,ds \right\|_1 \tag{3.27}$$
$$\leq \int_0^t \|\nabla f_0(\cdot,t,s)\|_1 ds \leq t\,\|f_0\|_{V(J_T^2;\Omega)}.$$

By Proposition 3.2.6, since $F_1(\mathbf{x},t) = \exp(\int_0^t \hat{f}_1(\mathbf{x},s)\,ds)$ with $f_1 := q^- - \nabla\cdot\mathbf{u} \in W^{1,\infty}(\Omega)$ is an exponential,

$$\begin{aligned} |(F_1 c^0)(\cdot,t)|_{BV} &\leq \|F_1(\cdot,t)\|_\infty |c^0|_{BV} + \|\nabla F_1(\cdot,t)\|_\infty \|c^0\|_1 \\ &\leq \|F_1(\cdot,t)\|_\infty \left( |c^0|_{BV} + \left\| \nabla \int_0^t \hat{f}_1(\cdot,s)\,ds \right\|_\infty \|c^0\|_1 \right) \\ &\leq \exp\big(t\|f_1\|_\infty\big)\big(|c^0|_{BV} + t\|\nabla f_1\|_\infty \|\nabla \hat{\mathbf{x}}\|_\infty \|c^0\|_1\big). \end{aligned} \tag{3.28}$$

Substituting (3.27) and (3.28) into (3.26) gives

$$|\hat{c}(\cdot,t)|_{BV} \leq C_0 t + e^{C_1 t}|c^0|_{BV}. \tag{3.29}$$

Noticing that $c = \hat{c}\circ\check{\mathbf{x}}$, and $\nabla\check{\mathbf{x}}(\cdot,0) \equiv I$ by (3.12), we have, by Proposition 3.2.7,

$$|c(\cdot,t)|_{BV} \leq (1+Ct)|\hat{c}(\cdot,t)|_{BV} \tag{3.30}$$

for some constant $C > 0$. Combining (3.29) and (3.30) gives (3.25) and completes the proof. □

For a general mesh $\mathcal{T}_h$ in multidimensional spaces, the $L^2$-projection operator $P_h$ might increase the variation of a function, so we cannot expect the TVB property to hold for the numerical solution. Instead, we make a weaker assumption of $L^1$-TVB as follows.

*Assumption* 3.2.5 ($L^1$-TVB of the numerical solution). The numerical solution $c_h^n$ to the system is uniformly $L^1$-TVB, i.e., there exists a constant $M > 0$ such that

$$|c_h^n|_{L^1(J_T;BV)} := \sum_{n=0}^{N-1} |c_h^n|_{BV}\Delta t^n \leq M$$

for any $h$, $\Delta t > 0$.

In particular, for rectangular meshes, we show next that the $L^2$-projection operator $P_h$ is TVD, so the numerical solution is TVB and $L^1$-TVB.

**Lemma 3.2.10.** *Let $\mathcal{T}_h$ be a rectangular mesh of a rectangular domain $\Omega = \prod_{i=1}^{d}(a_i, b_i) \subset \mathbb{R}^d$. Then $P_h$ is TVD, i.e., for any $f \in BV(\Omega)$,*

$$|P_h f|_{BV} \leq |f|_{BV}. \tag{3.31}$$

*Proof.* Let each interval $(a_i, b_i)$, $1 \leq i \leq d$, be partitioned into

$$a_i = x_i^0 < x_i^1 < \cdots < x_i^{n_i} = b_i$$

and each subinterval $I_i^j := (x_i^{j-1}, x_i^j)$ has length $h_i^j = x_i^j - x_i^{j-1}$, $1 \leq j \leq n_i$. Define the rectangular mesh $\mathcal{T}_h = \{E_{\mathbf{j}}\}_{\mathbf{j}\in\mathcal{J}}$, where the set of multi-indices $\mathcal{J}$ is

$$\mathcal{J} := \{\mathbf{j} = (j_1, j_2, \cdots, j_d) \in \mathbb{N}^d : 1 \leq j_i \leq n_i, 1 \leq i \leq d\}$$

and the grid element $E_{\mathbf{j}} := \prod_{i=1}^{d} I_i^{j_i}$.

By Propositions 3.2.4 and 3.2.5, we only need to show (3.31) for $f \in C^\infty(\Omega)$. Define a function of a single variable $x_i \in (a_i, b_i)$ to be

$$F_{\mathbf{j}}^i(x_i) := \int_{I_1^{j_1}} \cdots \int_{I_{i-1}^{j_{i-1}}} \int_{I_{i+1}^{j_{i+1}}} \cdots \int_{I_d^{j_d}} f(\mathbf{x})\, dx_d \cdots dx_{i+1}\, dx_{i-1} \cdots dx_1.$$

Now $P_h f \in W_h(\Omega)$ is piecewise constant, so its variation can be computed as the sum of each jump of $P_h f$ across an interface of adjacent elements multiplied by the projection area of the corresponding interface in each direction of the standard unit vector $\mathbf{e}_i$. For a pair of adjacent elements $E_{\mathbf{j}}$ and $E_{\mathbf{j}+\mathbf{e}_i}$, $1 \leq i \leq d$, $1 \leq j \leq n_i - 1$, the jump of $P_h f$ multiplied by the projection area of the corresponding interface

is

$$\begin{aligned}
&\left|\frac{1}{|E_{\mathbf{j}}|}\int_{E_{\mathbf{j}}} f(\mathbf{x})d\mathbf{x} - \frac{1}{|E_{\mathbf{j}+\mathbf{e}_i}|}\int_{E_{\mathbf{j}+\mathbf{e}_i}} f(\mathbf{x})d\mathbf{x}\right| \prod_{k\neq i} h_k^{j_k} \\
&\quad = \left|\frac{1}{h_i^{j_i}}\int_{I_i^{j_i}} F_{\mathbf{j}}^i(x_i)\,dx_i - \frac{1}{h_i^{j_i+1}}\int_{I_i^{j_i+1}} F_{\mathbf{j}}^i(x_i)\,dx_i\right| \\
&\quad = \left|F_{\mathbf{j}}^i(\xi_i^{j_i}) - F_{\mathbf{j}}^i(\xi_i^{j_i+1})\right| \\
&\quad \le \int_{\xi_i^{j_i}}^{\xi_i^{j_i+1}} \left|(F_{\mathbf{j}}^i)'(x_i)\right| dx_i,
\end{aligned}$$

where $\xi_i^{j_i} \in I_i^{j_i}$ and $\xi_i^{j_i+1} \in I_i^{j_i+1}$ are some points such that mean values of $F_{\mathbf{j}}^i$ in $I_i^{j_i}$ and $I_i^{j_i+1}$ are achieved, respectively. Then the total variation

$$\begin{aligned}
|P_h f|_{BV} &\le \sum_{i=1}^{d} \sum_{\mathbf{j}\in\mathfrak{I}, j_i<n_i} \int_{\xi_i^{j_i}}^{\xi_i^{j_i+1}} |(F_{\mathbf{j}}^i)'(x_i)|dx_i \\
&\le \sum_{i=1}^{d} \sum_{j_k=1,k\neq i}^{n_k} \int_{a_i}^{b_i} |(F_{\mathbf{j}}^i)'(x_i)|dx_i \\
&\le \sum_{i=1}^{d} \sum_{j_k=1,k\neq i}^{n_k} \sum_{j_i=1,}^{n_i} \|\partial_i f\|_{1,E_{\mathbf{j}}} \\
&= \sum_{i=1}^{d} \|\partial_i f\|_1 = \|\nabla f\|_1 = |f|_{BV}.
\end{aligned}$$

So we obtain (3.31) and complete the proof. □

## 3.3 An Approximation of Errors in the $L^1$-norm

In this section, we introduce an approximation of errors in the $L^1$-norm that plays an important role later in the convergence proof in Section 3.4. This approximation was first introduced by Kuznetsov [40] in the error estimates of conservation law (3.18)–(3.19) by the smoothing method and the viscosity method. It was later used by Lucier [42] in the error estimates of Glimm's method and Godunov's method.

Without losing generality, we assume $0 \in \Omega$. Let $K_\varepsilon$ be an approximation of the identity in $\Omega$, i.e.,

$$K_\varepsilon(\mathbf{x}) := \frac{1}{\varepsilon^d} K_0\left(\frac{\mathbf{x}}{\varepsilon}\right), \qquad \varepsilon > 0,$$

where function $K_0$ is non-negative, smooth and compactly supported in $\Omega$ with an integral of one. For the weak solution $c$ and the numerical solution $c_h$, we introduce

$$\rho^n_{\varepsilon,h} := \iint_{\Omega\times\Omega} K_\varepsilon(\mathbf{x}-\mathbf{y})|c^n(\mathbf{x}) - c^n_h(\mathbf{y})|\, d\mathbf{x}\, d\mathbf{y}. \tag{3.32}$$

Since $c^n_h = P_h c^{n-}_h$, we also have $\rho^{n-}_{\varepsilon,h}$ defined with $c^n_h$ replaced by $c^{n-}_h$.

**Lemma 3.3.1.** *The quantity $\rho^n_{\varepsilon,h}$ is an approximation of the $L^1$-error with a first order convergence rate with respect to $\varepsilon$. That is,*

$$\left|\, \rho^n_{\varepsilon,h} - \|c^n - c^n_h\|_1 \,\right| \le C\varepsilon, \tag{3.33}$$

*where $C > 0$ is a constant independent of $\varepsilon$, $h$, and $n$.*

*Proof.* For any fixed $\mathbf{y} \in \Omega$, when $\varepsilon$ is sufficiently small, $\Omega \subset \varepsilon^{-1}(\Omega - \{\mathbf{y}\})$, so

$$\int_\Omega K_\varepsilon(\mathbf{x}-\mathbf{y})\, d\mathbf{x} = \int_{\varepsilon^{-1}(\Omega-\{\mathbf{y}\})} K_0(\mathbf{x})\, d\mathbf{x} = \int_\Omega K_0(\mathbf{x})\, d\mathbf{x} = 1,$$

and so, by Proposition 3.2.8 and Lemma 3.2.9

$$\begin{aligned}
\left|\, \rho^n_{\varepsilon,h} - \|c^n - c^n_h\|_1 \,\right| &= \left| \iint_{\Omega\times\Omega} K_\varepsilon(\mathbf{x}-\mathbf{y})\left(|c^n(\mathbf{x}) - c^n_h(\mathbf{y})| - |c^n(\mathbf{y}) - c^n_h(\mathbf{y})|\right) d\mathbf{x}\, d\mathbf{y} \right| \\
&\le \iint_{\Omega\times\Omega} K_\varepsilon(\mathbf{x}-\mathbf{y})|c^n(\mathbf{x}) - c^n(\mathbf{y})|\, d\mathbf{x}\, d\mathbf{y} \\
&= \int_\Omega K_0(\mathbf{x}) \int_{\Omega_{\varepsilon\mathbf{x}}} |c^n(\mathbf{y}+\varepsilon\mathbf{x}) - c^n(\mathbf{y})|\, d\mathbf{y}\, d\mathbf{x} \\
&\le \int_\Omega K_0(\mathbf{x})\, d\mathbf{x}\ \varepsilon h_\Omega\ |c^n|_{BV(\Omega)} \le C\varepsilon,
\end{aligned}$$

where $h_\Omega$ is the diameter of domain $\Omega$. ∎

By changing variables, the definition (3.32) of $\rho^n_{\varepsilon,h}$ can be rewritten as

$$\rho^n_{\varepsilon,h} = \int_\Omega K_0(\mathbf{x}) \|T_{\varepsilon\mathbf{x}} c^n - c^n_h\|_{1,\Omega_{\varepsilon\mathbf{x}}} d\mathbf{x}.$$

Then we can again employ the entropy inequality (3.17) to prove the following lemma, which gives the estimate of the change of $\rho^n_{\varepsilon,h}$ in time.

**Lemma 3.3.2.** *The change of $\rho^n_{\varepsilon,h}$ in a single time step $J^n = [t^n, t^{n+1})$ has the estimate*

$$\rho^{n+1-}_{\varepsilon,h} - \rho^n_{\varepsilon,h} \le C(\varepsilon + h + (\Delta t)^r)\Delta t^n, \tag{3.34}$$

*where $r$ is given in* (3.6) *and $C > 0$ is a constant independent of $\varepsilon$, $h$, $\Delta t^n$, and $n$.*

*Proof.* Let $\varepsilon > 0$ and $\mathbf{x} \in \Omega$ fixed. Notice that from (2.8) and (3.7) the translated difference $d_{\varepsilon\mathbf{x},h} := T_{\varepsilon\mathbf{x}} c - c_h$ solves the linear balance law

$$(d_{\varepsilon\mathbf{x},h})_t + \nabla \cdot (d_{\varepsilon\mathbf{x},h}\tilde{\mathbf{u}}) = d_{\varepsilon\mathbf{x},h} q^- + R_{\varepsilon\mathbf{x}} \quad \text{in } \Omega_{\varepsilon\mathbf{x}} \times J^n,$$

where the reminder

$$R_{\mathbf{x}} := \nabla \cdot \big((T_{\mathbf{x}} c)(\tilde{\mathbf{u}} - T_{\mathbf{x}}\mathbf{u})\big) + \big(T_{\mathbf{x}}(c_I q^+) - c_I q^+\big) + (T_{\mathbf{x}} c)(T_{\mathbf{x}} q^- - q^-) \text{ for } \mathbf{x} \in \mathbb{R}^d.$$

For entropy $\eta(d) = |d|$ and test function $\varphi(\mathbf{x}, t) \equiv 1$, the entropy inequality (3.17) is reduced to

$$\begin{aligned} &\|d^n_{\varepsilon\mathbf{x},h}\|_{1,\Omega_{\varepsilon\mathbf{x}}} - \|d^{n+1-}_{\varepsilon\mathbf{x},h}\|_{1,\Omega_{\varepsilon\mathbf{x}}} \\ &\quad - \int_{J^n}\int_{\partial\Omega_{\varepsilon\mathbf{x}}} |d_{\varepsilon\mathbf{x},h}|\, \tilde{\mathbf{u}} \cdot \boldsymbol{\nu}\, ds\, dt + \int_{J^n} \|R_{\varepsilon\mathbf{x}}\|_{1,\Omega_{\varepsilon\mathbf{x}}} dt \ge 0, \end{aligned} \tag{3.35}$$

where, with Lemmas 3.2.1 and 3.2.2, Assumption 3.0.1, and (2.9),

$$\begin{aligned}
\int_{\partial\Omega_{\varepsilon\mathbf{x}}} |d_{\varepsilon\mathbf{x},h}|\, \tilde{\mathbf{u}}\cdot\boldsymbol{\nu}\, ds &\le |\partial\Omega|\, \|d_{\varepsilon\mathbf{x},h}\|_\infty \|\tilde{\mathbf{u}}\cdot\boldsymbol{\nu}\|_{\infty,\partial\Omega_{\varepsilon\mathbf{x}}} \\
&\le |\partial\Omega|\, \|d_{\varepsilon\mathbf{x},h}\|_\infty (\|\mathbf{u}\cdot\boldsymbol{\nu}\|_{\infty,\partial\Omega_{\varepsilon\mathbf{x}}} + \|(\tilde{\mathbf{u}}-\mathbf{u})\cdot\boldsymbol{\nu}\|_{\infty,\partial\Omega_{\varepsilon\mathbf{x}}}) \\
&\le |\partial\Omega|\, \|d_{\varepsilon\mathbf{x},h}\|_\infty (\|(T_{\varepsilon\mathbf{x}}\mathbf{u})\cdot\boldsymbol{\nu}\|_{\infty,\partial\Omega} + \|\tilde{\mathbf{u}}-\mathbf{u}\|_\infty) \\
&= |\partial\Omega|\, \|d_{\varepsilon\mathbf{x},h}\|_\infty (\|(T_{\varepsilon\mathbf{x}}\mathbf{u}-\mathbf{u})\cdot\boldsymbol{\nu}\|_{\infty,\partial\Omega} + \|\tilde{\mathbf{u}}-\mathbf{u}\|_\infty) \\
&\le |\partial\Omega|(\|c\|_\infty + \|c_h\|_\infty)\,(\varepsilon h_\Omega \|\nabla\mathbf{u}\|_\infty + C'(h+(\Delta t)^r)) \\
&\le C\,(\varepsilon + h + (\Delta t)^r)
\end{aligned}$$

and, with also Assumption 3.0.1 and Propositions 3.2.6 and 3.2.8, for any $t \in J^n$,

$$\begin{aligned}
\|R_{\varepsilon\mathbf{x}}\|_{1,\Omega_{\varepsilon\mathbf{x}}} &\le \|\nabla(T_{\varepsilon\mathbf{x}}c)\|_{1,\Omega_{\varepsilon\mathbf{x}}} \|\tilde{\mathbf{u}} - T_{\varepsilon\mathbf{x}}\mathbf{u}\|_\infty + \|T_{\varepsilon\mathbf{x}}c\|_{1,\Omega_{\varepsilon\mathbf{x}}} \|\nabla\cdot\tilde{\mathbf{u}} - T_{\varepsilon\mathbf{x}}\nabla\cdot\mathbf{u}\|_\infty \\
&\quad + \|T_{\varepsilon\mathbf{x}}(c_I q^+) - c_I q^+\|_{1,\Omega_{\varepsilon\mathbf{x}}} + \|(T_{\varepsilon\mathbf{x}}c)(T_{\varepsilon\mathbf{x}}q^- - q^-)\|_{1,\Omega_{\varepsilon\mathbf{x}}} \\
&\le \|\nabla c\|_1 (\|\tilde{\mathbf{u}} - \mathbf{u}\|_\infty + \|\mathbf{u} - T_{\varepsilon\mathbf{x}}\mathbf{u}\|_\infty) \\
&\quad + \|c\|_1 (\|\nabla\cdot\tilde{\mathbf{u}} - \nabla\cdot\mathbf{u}\|_\infty + \|\nabla\cdot\mathbf{u} - T_{\varepsilon\mathbf{x}}\nabla\cdot\mathbf{u}\|_\infty) \\
&\quad + \varepsilon h_\Omega (|c_I q^+|_{BV} + \|c\|_\infty |q|_{BV}) \\
&\le \|\nabla c\|_1 \left[C'(h+(\Delta t)^r) + \varepsilon h_\Omega \|\nabla\mathbf{u}\|_\infty\right] + \|c\|_1 \left[C'(h+(\Delta t)^r) + \varepsilon h_\Omega L\right] \\
&\quad + \varepsilon h_\Omega (|c_I|_{BV}\|q\|_\infty + \|c_I\|_1 \|\nabla q\|_\infty + \|c\|_\infty |q|_{BV}) \\
&\le C\,(\varepsilon + h + (\Delta t)^r)\,.
\end{aligned}$$

So (3.35) will be

$$\|d^n_{\varepsilon\mathbf{x},h}\|_{1,\Omega_{\varepsilon\mathbf{x}}} - \|d^{n+1-}_{\varepsilon\mathbf{x},h}\|_{1,\Omega_{\varepsilon\mathbf{x}}} + C(\varepsilon + h + (\Delta t)^r)\Delta t^n \ge 0$$

for some constant $C > 0$. Multiplying by $K_0(\mathbf{x})$ and integrating with respect to $\mathbf{x} \in \Omega$, we obtain (3.34) and complete the proof. □

The following lemma gives an estimate of the projection error measured by $\rho^n_{\varepsilon,h}$.

**Lemma 3.3.3.** *The projection error has the estimate*

$$\rho^n_{\varepsilon,h} - \rho^{n-}_{\varepsilon,h} \le C\frac{h^2}{\varepsilon}|c^{n-}_h|_{BV(\Omega)}, \tag{3.36}$$

*where $C > 0$ is a constant independent of $\varepsilon$, $h$, and $n$.*

*Proof.* We compute

$$\begin{aligned}
&\rho^n_{\varepsilon,h} - \rho^{n-}_{\varepsilon,h} \\
&= \iint_{\Omega\times\Omega} K_\varepsilon(\mathbf{x}-\mathbf{y})\{|c^n(\mathbf{x}) - P_h c^{n-}_h(\mathbf{y})| - |c^n(\mathbf{x}) - c^{n-}_h(\mathbf{y})|\}\,d\mathbf{x}d\mathbf{y} \\
&= \int_\Omega \sum_{E\in\mathcal{T}_h}\int_E K_\varepsilon(\mathbf{x}-\mathbf{y})\left\{\left|c^n(\mathbf{x}) - \frac{1}{|E|}\int_E c^{n-}_h(\mathbf{z})\,d\mathbf{z}\right| - \left|c^n(\mathbf{x}) - c^{n-}_h(\mathbf{y})\right|\right\}d\mathbf{x}d\mathbf{y} \\
&\le \int_\Omega \sum_{E\in\mathcal{T}_h}\frac{1}{|E|}\iint_{E\times E} K_\varepsilon(\mathbf{x}-\mathbf{y})\{|c^n(\mathbf{x}) - c^{n-}_h(\mathbf{z})| - |c^n(\mathbf{x}) - c^{n-}_h(\mathbf{y})|\}\,d\mathbf{z}d\mathbf{y}\,d\mathbf{x}.
\end{aligned}$$

If we switch variables $\mathbf{y}$ and $\mathbf{z}$ in the last inequality, the value simply changes sign, so the inequality can be written as

$$\begin{aligned}
&\rho^n_{\varepsilon,h} - \rho^{n-}_{\varepsilon,h} \\
&\le \frac{1}{2}\int_\Omega \sum_{E\in\mathcal{T}_h}\frac{1}{|E|}\iint_{E\times E}\{K_\varepsilon(\mathbf{x}-\mathbf{y}) - K_\varepsilon(\mathbf{x}-\mathbf{z})\} \\
&\qquad\qquad \times\{|c^n(\mathbf{x}) - c^{n-}_h(\mathbf{z})| - |c^n(\mathbf{x}) - c^{n-}_h(\mathbf{y})|\}\,d\mathbf{z}d\mathbf{y}\,d\mathbf{x} \\
&\le \frac{1}{2}\int_\Omega \sum_{E\in\mathcal{T}_h}\frac{1}{|E|}\iint_{E\times E}|K_\varepsilon(\mathbf{x}-\mathbf{y}) - K_\varepsilon(\mathbf{x}-\mathbf{z})||c^{n-}_h(\mathbf{z}) - c^{n-}_h(\mathbf{y})|\,d\mathbf{z}d\mathbf{y}\,d\mathbf{x} \\
&= \frac{1}{2}\sum_{E\in\mathcal{T}_h}\frac{1}{|E|}\iint_{E\times E}\left(\int_\Omega |K_\varepsilon(\mathbf{x}-\mathbf{y}) - K_\varepsilon(\mathbf{x}-\mathbf{z})|\,d\mathbf{x}\right)|c^{n-}_h(\mathbf{z}) - c^{n-}_h(\mathbf{y})|\,d\mathbf{z}d\mathbf{y}.
\end{aligned} \tag{3.37}$$

For any $\mathbf{y},\mathbf{z}\in E$, we have by Proposition 3.2.8 that

$$\begin{aligned}
\int_\Omega |K_\varepsilon(\mathbf{x}-\mathbf{y}) - K_\varepsilon(\mathbf{x}-\mathbf{z})|\,d\mathbf{x} &= \int_{\varepsilon^{-1}(\Omega-\{\mathbf{y}\})}\left|K_0(\mathbf{x}) - K_0\left(\mathbf{x}+\frac{\mathbf{y}-\mathbf{z}}{\varepsilon}\right)\right|d\mathbf{x} \\
&\le \frac{|\mathbf{y}-\mathbf{z}|}{\varepsilon}|K_0|_{BV(\Omega)} \le \frac{h}{\varepsilon}|K_0|_{BV(\Omega)},
\end{aligned} \tag{3.38}$$

and

$$\iint_{E\times E} |c_h^{n-}(\mathbf{z}) - c_h^{n-}(\mathbf{y})| d\mathbf{z} d\mathbf{y} = \int_E \left( \int_{E-\{\mathbf{z}\}} |c_h^{n-}(\mathbf{z}) - c_h^{n-}(\mathbf{y}+\mathbf{z})| d\mathbf{y} \right) d\mathbf{z} \quad (3.39)$$
$$\leq \int_{E-E} \left( \int_{E_{\mathbf{y}}} |c_h^{n-}(\mathbf{z}) - c_h^{n-}(\mathbf{y}+\mathbf{z})| d\mathbf{z} \right) d\mathbf{y}$$
$$\leq h|E-E|\ |c_h^{n-}|_{BV(E)},$$

where $E - E := \{\mathbf{x} - \mathbf{y} : \mathbf{x}, \mathbf{y} \in E\}$.

Let $B_r$ be a ball in $\mathbb{R}^d$ with radius $r > 0$, then by regularity of $\mathcal{T}_h$ in Assumption 3.2.4, we have

$$\frac{|E-E|}{|E|} \leq \frac{|B_{h_E}|}{|B_{\rho_E/2}|} = \left(\frac{2h_E}{\rho_E}\right)^d \leq (2\lambda_2)^d \text{ for any } E \in \mathcal{T}_h.$$

Substituting (3.38) and (3.39) into (3.37), we have by Proposition 3.2.3 that

$$\rho_{\varepsilon,h}^n - \rho_{\varepsilon,h}^{n-} \leq \frac{1}{2} \sum_{E\in T_h} \frac{1}{|E|} \frac{h}{\varepsilon} |K_0|_{BV(\Omega)}\ h|E-E|\ |c_h^{n-}|_{BV(E)}$$
$$\leq \frac{h^2}{2\varepsilon} |K_0|_{BV(\Omega)}\ (2\lambda_2)^d\, |c_h^{n-}|_{BV(\Omega)} \leq C\frac{h^2}{\varepsilon} |c_h^{n-}|_{BV(\Omega)}.$$

□

## 3.4 Convergence Results

We are ready to prove the following theorem on the convergence rate based on the previous lemmas.

**Theorem 3.4.1** (Convergence of VCCMM)**.** *Let Assumptions* 3.0.1 *and* 3.2.1–3.2.5 *hold* (*or omit Assumption* 3.2.5 *and assume* $\mathcal{T}_h$ *is rectangular*)*. Then the following* $L^1$*-error estimate holds:*

$$\max_{0\leq n\leq N} \|c_h^n - c^n\|_1 \leq \|c_h^0 - c^0\|_1 + C\left(\frac{h}{\sqrt{\Delta t}} + h + (\Delta t)^r\right), \quad (3.40)$$

*where* $C > 0$ *is a constant independent of* $h$ *and* $\Delta t$*.*

*Proof.* Summing (3.34) for $n$ in Lemma 3.3.2, we have

$$\sum_{k=0}^{n-1}(\rho_{\varepsilon,h}^{k+1-}-\rho_{\varepsilon,h}^{k})\leq C(\varepsilon+h+(\Delta t)^{r})t^{n}\leq CT(\varepsilon+h+(\Delta t)^{r}).$$

Rearranging, we have

$$\rho_{\varepsilon,h}^{n}\leq\rho_{\varepsilon,h}^{0}+E_{\varepsilon,h}^{n}+CT(\varepsilon+h+(\Delta t)^{r}),\tag{3.41}$$

where the total projection error is

$$E_{\varepsilon,h}^{n}:=\sum_{k=1}^{n}(\rho_{\varepsilon,h}^{k}-\rho_{\varepsilon,h}^{k-}).$$

Summing (3.36) for $n$ in Lemma 3.3.3, we have

$$E_{\varepsilon,h}^{n}\leq C\frac{h^{2}}{\varepsilon}\sum_{k=1}^{n}|c_{h}^{k-}|_{BV}.\tag{3.42}$$

By (3.25) in time step $J^{k-1}$,

$$|c_{h}^{k-}|_{BV}\leq C_{0}\Delta t^{k-1}+e^{C_{1}\Delta t^{k-1}}|c_{h}^{k-1}|_{BV},$$

so substituting into (3.42) gives

$$\begin{aligned}E_{\varepsilon,h}^{n}&\leq C\frac{h^{2}}{\varepsilon}\left(C_{0}T+e^{C_{1}T}\sum_{k=1}^{n}|c_{h}^{k-1}|_{BV}\right)\\&\leq C\frac{h^{2}}{\varepsilon\Delta t}\left(C_{0}T^{2}+\lambda_{1}e^{C_{1}T}\sum_{k=1}^{n}|c_{h}^{k-1}|_{BV}\Delta t^{k-1}\right)\\&\leq C\frac{h^{2}}{\varepsilon\Delta t}\left(C_{0}T^{2}+\lambda_{1}e^{C_{1}T}|c_{h}^{k}|_{L^{1}(J_{T};BV)}\right)\\&\leq C\frac{h^{2}}{\varepsilon\Delta t}\left(C_{0}T^{2}+\lambda_{1}Me^{C_{1}T}\right),\end{aligned}\tag{3.43}$$

where $|c_{h}^{k}|_{L^{1}(J_{T};BV)}\leq M$ by Assumption 3.2.5. Combining estimates (3.41), (3.43), and (3.33) gives

$$\|c_{h}^{n}-c^{n}\|_{1}\leq\|c_{h}^{0}-c^{0}\|_{1}+C\left(\varepsilon+\frac{h^{2}}{\varepsilon\Delta t}+h+(\Delta t)^{r}\right),$$

where the optimal choice for $\varepsilon$ is to take $\varepsilon=h/\sqrt{\Delta t}$, which completes the proof. □

*Remark* 3.4.1. Note that there is no CFL constraint on $\Delta t$ in (3.40), so, theoretically, $\Delta t$ can be taken independent of $h$. The optimal choice is to take $\Delta t = Ch^{2/(2r+1)}$, which leads to the error $\mathcal{O}(h^{2r/(2r+1)})$. In practice, we take $\Delta t = Ch$ to avoid possibly generated self-intersected trace-back polygons, which leads to the error $\mathcal{O}(h^{1/2})$.

## 3.5 The Existence of the Perturbed Velocity

In this section, we make several assumptions that will guarantee the existence of the perturbed velocity field $\tilde{\mathbf{u}}$ satisfying the requirements of Assumption 3.0.1. That is, we prove Assumption 3.0.1 by constructing a perturbed velocity field $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}(\mathbf{x}, t)$ on the domain $\Omega \times J_T$. We need to impose assumptions on the choices of rings that are adjusted in Step 2 of the Volume Correction Algorithm in Section 2.2. For simplicity, we concentrate on the case that the domain $\Omega \subset \mathbb{R}^2$, although the ideas can carry over to higher spatial dimensions. Below we consider the effect of three main steps of volume adjustment: characteristic time perturbation, ring adjustment, and individual element adjustment. Note that, for ease of exposition, we do not treat forward tracing around wells, though clearly the ideas of the proof extend to this step.

*Remark* 3.5.1. In the rest of this section, we tacitly assume that the velocity field $\mathbf{u}$ is given by a quarter of a "five-spot" pattern of wells, which is a rectangular domain with an injection well near a corner and a production well near the opposite corner.

### 3.5.1 Point adjustment in time and the local definition of $\tilde{\mathbf{u}}$

The following lemma constructs a perturbed velocity locally at isolated points and quantifies how a small trace-back time perturbation of size $\alpha \Delta t^n$ changes a single characteristic trace-back.

**Lemma 3.5.1.** *Suppose $\alpha \in \mathbb{R}$ is fixed and $\mathbf{x} \in \overline{\Omega}$. For $t \in J^n$, let*

$$\tilde{\mathbf{x}}(t) = \tilde{\mathbf{x}}(\mathbf{x}, t) := \check{\mathbf{x}}\big(\mathbf{x}, t + \alpha(t^{n+1} - t)\big) \tag{3.44}$$

*be a time perturbation of the trace-back curve $\check{\mathbf{x}}(t)$. Then the perturbed velocity*

$$\tilde{\mathbf{u}}(\mathbf{x}, t) := (1 - \alpha)\, \mathbf{u}\big(\mathbf{x}, t + \alpha(t^{n+1} - t)\big) \tag{3.45}$$

*has $\tilde{\mathbf{x}}$ as its characteristic passing through point $\mathbf{x}$ at time $t^{n+1}$. Moreover,*

$$\tilde{\mathbf{u}} \cdot \boldsymbol{\nu} = 0 \quad \text{on } \partial\Omega \times J^n, \tag{3.46}$$

$$\|\mathbf{u} - \tilde{\mathbf{u}}\|_\infty + \|\nabla \cdot \mathbf{u} - \nabla \cdot \tilde{\mathbf{u}}\|_\infty \le C|\alpha|, \tag{3.47}$$

*where $C = (\|\mathbf{u}_t\|_\infty + \|\nabla \cdot \mathbf{u}_t\|_\infty)T + \|\mathbf{u}\|_\infty + \|\nabla \cdot \mathbf{u}\|_\infty$.*

*Proof.* We compute

$$\begin{aligned}
\tilde{\mathbf{x}}'(t) &= (1 - \alpha)\, \check{\mathbf{x}}'\big(t + \alpha(t^{n+1} - t)\big) \\
&= (1 - \alpha)\, \mathbf{u}\big(\check{\mathbf{x}}(t + \alpha(t^{n+1} - t)), t + \alpha(t^{n+1} - t)\big) \\
&= (1 - \alpha)\, \mathbf{u}\big(\tilde{\mathbf{x}}(t), t + \alpha(t^{n+1} - t)\big) = \tilde{\mathbf{u}}(\tilde{\mathbf{x}}(t), t),
\end{aligned}$$

and since clearly $\tilde{\mathbf{x}}(t^{n+1}) = \check{\mathbf{x}}(t^{n+1}) = \mathbf{x}$, we have the claimed characteristic curve. Now,

$$\begin{aligned}
|\mathbf{u}(\mathbf{x}, t) - \tilde{\mathbf{u}}(\mathbf{x}, t)| &= \big|\mathbf{u}(\mathbf{x}, t) - (1 - \alpha)\mathbf{u}\big(\mathbf{x}, t + \alpha(t^{n+1} - t)\big)\big| \\
&\le \big|\mathbf{u}(\mathbf{x}, t) - \mathbf{u}\big(\mathbf{x}, t + \alpha(t^{n+1} - t)\big)\big| + |\alpha|\|\mathbf{u}\|_\infty \\
&\le \big\{\|\mathbf{u}_t\|_\infty \Delta t^n + \|\mathbf{u}\|_\infty\big\}\, |\alpha| \le C|\alpha|,
\end{aligned}$$

and similarly for $|\nabla \cdot \mathbf{u}(\mathbf{x}, t) - \nabla \cdot \tilde{\mathbf{u}}(\mathbf{x}, t)|$, so (3.47) follows. By construction, the perturbed boundary condition (3.46) holds due to (2.9). □

*Remark* 3.5.2. In practice, the ordinary differential equation (2.5) for characteristics cannot be solved exactly unless the velocity field is particularly simple.

Therefore, numerical techniques are needed. For example, if the single Euler step is used, then we actually trace back from a point $\mathbf{x}_0$ with local velocity field

$$\mathbf{u}_{\mathrm{E}}(\mathbf{x}, t) := \mathbf{u}(\mathbf{x}_0, t^{n+1}),$$

where $\mathbf{x} = \mathbf{x}_0 - (t^{n+1} - t)\mathbf{u}(\mathbf{x}_0, t^{n+1})$ for $t \in J^n$. This leads to an error

$$\|\mathbf{u}_{\mathrm{E}} - \mathbf{u}\|_\infty + \|\nabla \cdot \mathbf{u}_{\mathrm{E}} - \nabla \cdot \mathbf{u}\|_\infty \leq C(\Delta t)^r,$$

where $r = 1$. Since $\mathbf{u} = \mathbf{u}_{\mathrm{E}} + (\mathbf{u} - \mathbf{u}_{\mathrm{E}})$, we simply replace $\mathbf{u}$ by $\mathbf{u}_{\mathrm{E}}$, and the rest of the analysis remains unchanged except that there is an extra error due to approximately solving for characteristics. In general, we may use an approximation of order $r > 1$. For ease of exposition, we tacitly omit this extra error term in this section.

*Remark* 3.5.3. If $\mathbf{u}$ is unknown, then we may need to approximate $\mathbf{u}$ with $\mathbf{u}_h$ by numerical techniques, which leads to an error $\varepsilon_{\text{flow}}$ due to flow approximation. If so, this error would enter the estimates as well. This case is handled in Chapter 7.

### 3.5.2 Global $\tilde{\mathbf{u}}$ and the ring adjustment

Now consider the ring adjustment phase of the Volume Correction Algorithm. We have defined a local perturbed velocity field $\tilde{\mathbf{u}}$ for a single characteristic in Lemma 3.5.1. Here we further perturb $\tilde{\mathbf{u}}$ to obtain volume conservation over rings of elements.

At time $t^{n+1}$, let $R \subset \Omega$ be a ring (Figure 3.1, left) and $\check{R}^n$ be the exact trace-back region with velocity field $\mathbf{u}$ for time $\Delta t^n$. Vertices and midpoints $\mathbf{x}_i$ on $\partial R$ are traced back to $\check{\mathbf{x}}_i^n$, where $1 \leq i \leq N_R$. Without losing generality, assume points $\check{\mathbf{x}}_i^n$, where $1 \leq i \leq N_{\text{ext}}$ for some $N_{\text{ext}} < N_R$, are on the "exterior" boundary (i.e., away from injection sites) of $\check{R}^n$ which need to be adjusted. We perturb these points in time of size $\alpha\Delta t^n$ as defined in Lemma 3.5.1, i.e., $\tilde{\mathbf{x}}_i^n = \check{\mathbf{x}}(\mathbf{x}_i, t^n + \alpha\Delta t^n)$, $1 \leq i \leq N_{\text{ext}}$. Denote this perturbed trace-back polygon as $\tilde{R}^n(\alpha)$ (Figure 3.1,

middle) with the "exterior" boundary $\Gamma^n(\alpha)$. We will choose the ring such that the shape of the ring is approximately "perpendicular" to the direction of the flow; that is, the volume change of the ring should be sensitive to the adjustment in the direction of the flow.

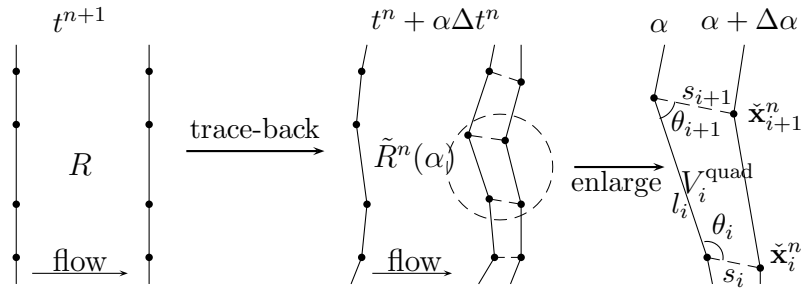


Figure 3.1: Ring $R$ at time $t^{n+1}$ is traced back to time $t^n$ and approximated by $\tilde{R}^n$. The solid dots represent the points which are traced back. The exterior boundary of $\tilde{R}^n$ is perturbed in location by a time change of $\alpha\Delta t^n$ and $(\alpha + \Delta\alpha)\Delta t^n$ along the direction of characteristics.

*Assumption* 3.5.1. There exists a constant $C' > 0$ independent of $h$ such that the number of vertices and midpoints on $\partial R$ satisfies $N_R \leq C' h^{-1}$.

*Assumption* 3.5.2 (Monotonicity and differentiability). When $\alpha_1 \leq \alpha_2$, $\tilde{R}^n(\alpha_1) \subseteq \tilde{R}^n(\alpha_2)$, and the pore volume $V_n(\alpha) := |\tilde{R}^n(\alpha)|_\phi$ is differentiable with respect to $\alpha$.

*Assumption* 3.5.3 (Non-degeneracy). There exist constants $\phi_* > 0$, $u_* > 0$, and $\Gamma_* > 0$ such that $1 \geq \phi(\mathbf{x}) \geq \phi_*$ in $\Omega$, $|\mathbf{u}| \geq u_*$ in a sufficiently large neighborhood of $\Gamma^n(\alpha)$, and $|\Gamma^n(\alpha)| \geq \Gamma_*$.

*Assumption* 3.5.4 (Non-parallelism). There exists a constant $\nu_* > 0$ such that $\mathbf{u} \cdot \boldsymbol{\nu}_\alpha \geq \nu_* |\mathbf{u}|$ in a neighborhood of $\Gamma^n(\alpha)$, where $\boldsymbol{\nu}_\alpha$ is the unit outward normal vector with respect to $\Gamma^n(\alpha)$.

*Remark* 3.5.4. The condition $|\Gamma^n(\alpha)| \geq \Gamma_*$ in Assumption 3.5.3 implies that the trace-back procedure should only be performed away from injection wells, where

points do not trace into the well-bore and become arbitrarily close. Therefore, a trace-forward technique is used near injection wells in the Volume Correction Algorithm. We note also that $|\mathbf{u}| > u_*$ in Assumption 3.5.3 does not cover the case of velocity fields with stagnation points when $\Gamma^n(\alpha)$ is near the point. The "sufficiently large" condition is defined in the proof of Lemma 3.5.3 below (see (3.53) ).

The following lemma shows the existence of the perturbed velocity field $\tilde{\mathbf{u}}$ such that the trace-back region of a ring $R$ satisfies the *local volume constraint* (2.16) in the absence of source $q$.

**Lemma 3.5.2.** *Let $R \subset \Omega$ be a ring to be adjusted. If Assumptions* 3.5.1–3.5.4 *hold, then there exists some $\alpha^*$ such that*

$$V_n(\alpha^*) = |\check{R}^n|_\phi, \tag{3.48}$$

*where $|\alpha^*| \le Ch$ for some constant $C > 0$ independent of $n$, $h$, and $\Delta t$.*

To show Lemma 3.5.2, we need another lemma which simply says that the change rate of the pore volume $V_n(\alpha)$ is bounded away from zero during the ring adjustment.

**Lemma 3.5.3.** *If Assumptions* 3.5.1–3.5.4 *hold, then*

$$V_n'(\alpha) \ge \beta_* \Delta t^n, \tag{3.49}$$

*where $\beta_* > 0$ is a constant independent of $n$, $h$, and $\Delta t^n$.*

*Proof.* For a small $\Delta\alpha > 0$, by Assumptions 3.5.2 and 3.5.3, we have

$$V_n(\alpha+\Delta\alpha) - V_n(\alpha) = |\tilde{R}^n(\alpha+\Delta\alpha) \setminus \tilde{R}^n(\alpha)|_\phi \ge \phi_* |\tilde{R}^n(\alpha+\Delta\alpha) \setminus \tilde{R}^n(\alpha)|, \tag{3.50}$$

where the set $\tilde{R}^n(\alpha+\Delta\alpha) \setminus \tilde{R}^n(\alpha)$ can be decomposed as a union of quadrilaterals (Figure 3.1, middle).

As shown in Figure 3.1 (right), the volume of each quadrilateral $V_i^{\text{quad}}$ ($1 \le i \le N'_{\text{ext}}$, $N'_{\text{ext}} = N_{\text{ext}}$ if the ring $R$ does not intersect $\partial\Omega$ and $N'_{\text{ext}} = N_{\text{ext}} - 1$ otherwise) is

$$\begin{aligned} V_i^{\text{quad}} &= \frac{1}{2}(s_i \sin\theta_i + s_{i+1}\sin\theta_{i+1})(l_i - s_i\cos\theta_i - s_{i+1}\cos\theta_{i+1}) \\ &\quad + \frac{1}{2}s_i^2 \sin\theta_i\cos\theta_i + \frac{1}{2}s_{i+1}^2 \sin\theta_{i+1}\cos\theta_{i+1} \\ &= \frac{1}{2}[\, l_i s_i \sin\theta_i + l_i s_{i+1}\sin\theta_2 - s_i s_{i+1}\sin(\theta_i + \theta_{i+1})\,] \\ &\ge \frac{1}{2}(\, l_i s_i \sin\theta_i + l_i s_{i+1}\sin\theta_{i+1} - s_i s_{i+1}\,). \end{aligned} \tag{3.51}$$

Each displacement $s_i$ is

$$\begin{aligned} s_i &= \left|\check{\mathbf{x}}_i\big(t^n + (\alpha+\Delta\alpha)\Delta t^n\big) - \check{\mathbf{x}}_i(t^n + \alpha\Delta t^n)\right| \\ &= \left|\int_{t^n+\alpha\Delta t^n}^{t^n+(\alpha+\Delta\alpha)\Delta t^n} \mathbf{u}(\check{\mathbf{x}}_i(t),t)\,dt\right| \le \|\mathbf{u}\|_\infty \Delta\alpha\Delta t^n, \end{aligned} \tag{3.52}$$

and by Assumptions 3.5.3 and 3.5.4,

$$\begin{aligned} s_i &= \left|\int_{t^n+\alpha\Delta t^n}^{t^n+(\alpha+\Delta\alpha)\Delta t^n} \mathbf{u}(\check{\mathbf{x}}_i(t),t)\,dt\right| \\ &\ge \left|\int_{t^n+\alpha\Delta t^n}^{t^n+(\alpha+\Delta\alpha)\Delta t^n} \mathbf{u}(\check{\mathbf{x}}_i(t),t)\cdot\boldsymbol{\nu}_\alpha\,dt\right| \ge \nu_* u_* \Delta\alpha\Delta t^n. \end{aligned} \tag{3.53}$$

Substituting (3.52), (3.53), and each $\sin\theta_i \ge \nu_*$ by Assumption 3.5.4 into (3.51) gives

$$V_i^{\text{quad}} \ge \left(\nu_*^2 u_* l_i - \frac{1}{2}\|\mathbf{u}\|_\infty^2 \Delta\alpha\Delta t^n\right)\Delta\alpha\Delta t^n. \tag{3.54}$$

To obtain a lower bound of the difference $V_n(\alpha+\Delta\alpha) - V_n(\alpha)$ in (3.50), summing (3.54) for all $V_i^{\text{quad}}$ in the ring $\tilde{R}^n(\alpha)$, by Assumptions 3.5.1 and 3.5.3, we have

$$\begin{aligned} V_n(\alpha+\Delta\alpha) - V_n(\alpha) &\ge \phi_* \sum_i V_i^{\text{quad}} \\ &\ge \phi_*\left(\nu_*^2 u_* |\Gamma^n(\alpha)| - \frac{N_{\text{ext}}}{2}\|\mathbf{u}\|_\infty^2\Delta\alpha\Delta t^n\right)\Delta\alpha\Delta t^n \\ &\ge \phi_*\left(\nu_*^2 u_*\Gamma_* - \frac{C\Delta t^n}{2h}\|\mathbf{u}\|_\infty^2\Delta\alpha\right)\Delta\alpha\Delta t^n. \end{aligned} \tag{3.55}$$

Divide by $\Delta\alpha$ and let $\Delta\alpha \to 0$ in (3.55). We obtain (3.49) with $\beta_* = \phi_* \nu_*^2 u_* \Gamma_*$. $\square$

Now we are ready to prove Lemma 3.5.2.

*Proof.* (Lemma 3.5.2) For any $\alpha$ in a neighborhood of zero, consider the difference

$$\begin{aligned} V_n(\alpha) - |\check{R}^n|_\phi &= (V_n(\alpha) - V_n(0)) + (V_n(0) - |\check{R}^n|_\phi) \\ &= V_n'(\xi)\alpha + (|\tilde{R}^n(0)|_\phi - |\check{R}^n|_\phi), \end{aligned} \tag{3.56}$$

where $\xi = \xi(\alpha)$ comes from the mean value theorem. For the second term on the right hand side, since $0 \le \phi \le 1$,

$$\big||\tilde{R}^n(0)|_\phi - |\check{R}^n|_\phi\big| \le \big|(\tilde{R}^n(0) \setminus \check{R}^n) \cup (\check{R}^n \setminus \tilde{R}^n(0))\big|, \tag{3.57}$$

which is the discrepancy of volumes between $\tilde{R}^n(0)$ and $\check{R}^n$. This discrepancy is the sum of the discrepancies associated to each edge.

As illustrated in Figure 3.2, at time $t^{n+1}$, let $e_i$ $(1 \le i \le N_R)$ be an edge of $R$ with ends $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ $(\mathbf{x}_{N_R+1} = \mathbf{x}_1)$, which is traced back with velocity $\mathbf{u}$ to a curve $\check{e}_i(t)$ at time $t \in J^n$ with ends $\check{\mathbf{x}}_i(t) = \check{\mathbf{x}}(\mathbf{x}_i, t)$ and $\check{\mathbf{x}}_{i+1}(t) = \check{\mathbf{x}}(\mathbf{x}_{i+1}, t)$. Curve $\check{e}_i(t)$ is approximated by a line segment $\tilde{e}_i(t)$ by connecting $\check{\mathbf{x}}_i(t)$ and $\check{\mathbf{x}}_{i+1}(t)$. Let $\tilde{\mathbf{e}}_i(t) := \check{\mathbf{x}}_{i+1}(t) - \check{\mathbf{x}}_i(t)$. The local discrepancy $V_i^{\text{dis}}$ at time $t^n$ associated to edge $e_i$ is the net difference in area using the correct curve $\check{e}_i^n$ versus the segment $\tilde{e}_i^n$.

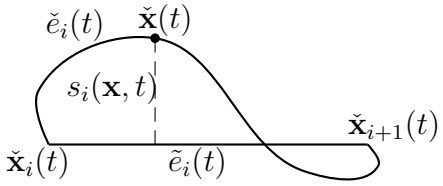


Figure 3.2: An edge $e$ of ring $R$ is traced back to a curve $\check{e}_i(t)$ with two ends $\check{\mathbf{x}}_i^n(t)$ and $\check{\mathbf{x}}_{i+1}^n(t)$, which is approximated by a line segment $\tilde{e}_i(t)$.

For any $\mathbf{x} \in e_i$ which is traced back to $\check{\mathbf{x}}^n(t) = \check{\mathbf{x}}(\mathbf{x}, t^n) \in \check{e}_i(t)$, let

$$s_i(\mathbf{x}, t) := \frac{\det\big(\check{\mathbf{x}}(t) - \check{\mathbf{x}}_i(t), \tilde{\mathbf{e}}_i(t)\big)}{|\tilde{\mathbf{e}}_i(t)|} \tag{3.58}$$

be the algebraic distance from point $\check{\mathbf{x}}(t)$ to segment $\tilde{e}_i(t)$, where $\det(\mathbf{x}, \mathbf{y})$ is the determinant of a $2 \times 2$ matrix formed by column vectors $\mathbf{x}$ and $\mathbf{y}$. Since $\check{\mathbf{x}}(\cdot, t)$ is a diffeomorphism by Assumption 3.2.1, $|\tilde{\mathbf{e}}_i(t)| \neq 0$, and (3.58) is well defined. Then

$$V_i^{\text{dis}} \leq 2\|s_i^n\|_{\infty, e_i} \sup_{\mathbf{x}, \mathbf{y} \in e_i} |\check{\mathbf{x}}^n - \check{\mathbf{y}}^n| \leq 2\|s_i^n\|_{\infty, e_i} \|\nabla \check{\mathbf{x}}\|_\infty h, \tag{3.59}$$

where $\|\nabla \check{\mathbf{x}}\|_\infty$ is bounded since, by taking gradients of (2.5) and (2.6), $\nabla \check{\mathbf{x}}$ solves the *linear* ordinary differential equation in time

$$\begin{aligned} (\nabla \check{\mathbf{x}})_t &= \nabla \mathbf{u}(\check{\mathbf{x}}, t) \nabla \check{\mathbf{x}} \quad && \text{in } \Omega \times J^n, \\ \nabla \check{\mathbf{x}}^{n+1} &= \mathbf{I} && \text{in } \Omega. \end{aligned}$$

At time $t^{n+1}$, $\check{\mathbf{x}}^{n+1} = \mathbf{x} \in e_i$, so by (3.58),

$$s_i^{n+1}(\mathbf{x}) = \frac{\det\big(\mathbf{x} - \mathbf{x}_i, \tilde{\mathbf{e}}_i^{n+1}\big)}{|\tilde{\mathbf{e}}_i^{n+1}|} = 0.$$

By the mean value theorem, there exists some $\tau \in J^n$ such that

$$\begin{aligned} \big|s_i^n(\mathbf{x})\big| &= \left|\frac{\partial s_i}{\partial t}(\mathbf{x}, \tau)\right| \Delta t^n \\ &= \left| \frac{\det\big(\check{\mathbf{x}}'(\tau) - \check{\mathbf{x}}_i'(\tau), \tilde{\mathbf{e}}_i(\tau)\big)}{|\tilde{\mathbf{e}}_i(\tau)|} + \frac{\det\big(\check{\mathbf{x}}(\tau) - \check{\mathbf{x}}_i(\tau), \tilde{\mathbf{e}}_i'(\tau)\big)}{|\tilde{\mathbf{e}}_i(\tau)|} \right. \\ &\qquad \left. - \det\big(\check{\mathbf{x}}(\tau) - \check{\mathbf{x}}_i(\tau), \tilde{\mathbf{e}}_i(\tau)\big) \frac{\tilde{\mathbf{e}}_i(\tau) \cdot \tilde{\mathbf{e}}_i'(\tau)}{|\tilde{\mathbf{e}}_i(\tau)|^3} \right| \Delta t^n. \end{aligned}$$

Applying inequalities $|\det(\mathbf{x}, \mathbf{y})| \leq |\mathbf{x}|\,|\mathbf{y}|$ for any $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^2$, and

$$|\tilde{\mathbf{e}}_i'(\tau)| = |\check{\mathbf{x}}_{i+1}'(\tau) - \check{\mathbf{x}}_i'(\tau)| = |\mathbf{u}(\check{\mathbf{x}}_{i+1}(\tau), \tau) - \mathbf{u}(\check{\mathbf{x}}_i(\tau), \tau)| \leq \|\nabla \mathbf{u}\|_\infty |\tilde{\mathbf{e}}_i(\tau)|,$$

we have

$$\begin{aligned}\left|s_i^n(\mathbf{x})\right| &\le \left(|\check{\mathbf{x}}'(\tau) - \check{\mathbf{x}}_i'(\tau)| + 2\|\nabla\mathbf{u}\|_\infty|\check{\mathbf{x}}(\tau) - \check{\mathbf{x}}_i(\tau)|\right)\Delta t^n \\ &= \left(|\mathbf{u}(\check{\mathbf{x}}(\tau),\tau) - \mathbf{u}(\check{\mathbf{x}}_i(\tau),\tau)| + 2\|\nabla\mathbf{u}\|_\infty|\check{\mathbf{x}}(\tau) - \check{\mathbf{x}}_i(\tau)|\right)\Delta t^n \\ &\le 3\|\nabla\mathbf{u}\|_\infty|\check{\mathbf{x}}(\tau) - \check{\mathbf{x}}_i(\tau)|\Delta t^n \\ &\le 3\|\nabla\mathbf{u}\|_\infty\|\nabla\check{\mathbf{x}}\|_\infty h\Delta t^n.\end{aligned} \tag{3.60}$$

Combining (3.59) and (3.60) gives

$$V_i^{\text{dis}} \le C''h^2\Delta t^n, \tag{3.61}$$

and summing over all edges $e_i$ of ring $\tilde{R}^n(0)$, by Assumption 3.5.1, we have

$$\left|(\tilde{R}^n(0) \setminus \check{R}^n) \cup (\check{R}^n \setminus \tilde{R}^n(0))\right| = \sum_{i=1}^{N_R} V_i^{\text{dis}} \le N_R C''h^2\Delta t^n \le C'C''h\Delta t^n. \tag{3.62}$$

Combining (3.56), (3.57), (3.62), and (3.49) gives

$$V_n(\alpha) - |\check{R}^n|_\phi \le C'C''h\Delta t^n + \beta_*\Delta t^n\alpha < 0 \text{ when } \alpha < -\frac{C'C''}{\beta_*}h, \tag{3.63}$$

$$V_n(\alpha) - |\check{R}^n|_\phi \ge -C'C''h\Delta t^n + \beta_*\Delta t^n\alpha > 0 \text{ when } \alpha > \frac{C'C''}{\beta_*}h. \tag{3.64}$$

By the continuity of $V_n(\alpha) - |\check{R}^n|_\phi$, inequalities (3.63) and (3.64) imply that there exists some $\alpha^*$, where $|\alpha^*| \le Ch$, such that equation (3.48) holds. □

### 3.5.3 Individual element adjustment

Finally, we consider the individual element adjustment of the Volume Correction Algorithm. Let $E$ is a grid element in a ring $R$, and $\mathbf{x}_{\text{m}}$ be the midpoint of an edge $e = \mathbf{x}_{\text{l}}\mathbf{x}_{\text{r}}$ of $E$ between the inner and outer ring boundaries which requires adjustment. Vertices and midpoints of edges of $E$ are traced back for time $\Delta t^n$ and are adjusted to a polygon $\tilde{E}^n(\alpha^*)$ (Figure 3.3, left) in the ring adjustment, where $\alpha^*$ is determined by (3.48). The following lemma gives the local construction of the perturbed velocity field near midpoint $\mathbf{x}_{\text{m}}$.

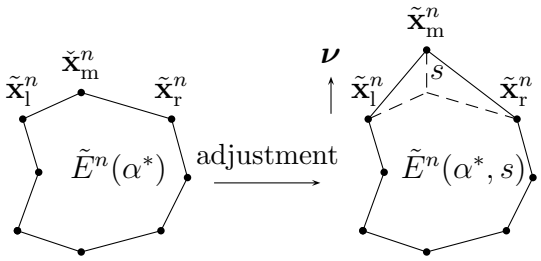


Figure 3.3: The trace-back midpoint $\check{\mathbf{x}}_{\mathrm{m}}^n$ of element $\tilde{E}^n(\alpha^*)$ is adjusted to $\tilde{\mathbf{x}}_{\mathrm{m}}^n$ in the direction of $\boldsymbol{\nu}$.

**Lemma 3.5.4.** *For $t \in J^n$, let*

$$\tilde{\mathbf{x}}(\mathbf{x}, t) := \check{\mathbf{x}}(\mathbf{x}, t) + s \left( \frac{t^{n+1} - t}{\Delta t^n} \right) \boldsymbol{\nu} \tag{3.65}$$

*be a perturbation of the trace-back characteristic $\check{\mathbf{x}}(t)$, so that, in particular, $\tilde{\mathbf{x}}_{\mathrm{m}}^n = \check{\mathbf{x}}_{\mathrm{m}}^n + s\boldsymbol{\nu}$ is a perturbation of the trace-back midpoint $\check{\mathbf{x}}_{\mathrm{m}}^n = \check{\mathbf{x}}(\mathbf{x}_{\mathrm{m}}, t^n)$, where $\boldsymbol{\nu}$ is the unit normal vector with respect to the trace-back segment $\tilde{\mathbf{x}}_{\mathrm{l}}^n \tilde{\mathbf{x}}_{\mathrm{r}}^n$, and $s \in \mathbb{R}$ is the adjustment distance (Figure 3.3, right). Then the perturbed velocity*

$$\tilde{\mathbf{u}}(\mathbf{x}, t) := \mathbf{u} \left( \mathbf{x} - s \left( \frac{t^{n+1} - t}{\Delta t^n} \right) \boldsymbol{\nu}, t \right) - \frac{s}{\Delta t^n} \boldsymbol{\nu} \tag{3.66}$$

*has $\tilde{\mathbf{x}}$ as its characteristic passing through point $\mathbf{x}$ at time $t^{n+1}$ and*

$$\|\mathbf{u} - \tilde{\mathbf{u}}\|_\infty + \|\nabla \cdot \mathbf{u} - \nabla \cdot \tilde{\mathbf{u}}\|_\infty \leq C \frac{|s|}{\Delta t^n}, \tag{3.67}$$

*where $C > 0$ is a constant independent of $h$ and $\Delta t^n$.*

*Proof.* We compute

$$\begin{aligned} \tilde{\mathbf{x}}'(t) &= \check{\mathbf{x}}'(t) - \frac{s}{\Delta t^n} \boldsymbol{\nu} = \mathbf{u}(\check{\mathbf{x}}(t), t) - \frac{s}{\Delta t^n} \boldsymbol{\nu} \\ &= \mathbf{u} \left( \tilde{\mathbf{x}}(t) - s \left( \frac{t^{n+1} - t}{\Delta t^n} \right) \boldsymbol{\nu}, t \right) - \frac{s}{\Delta t^n} \boldsymbol{\nu} = \tilde{\mathbf{u}}(\tilde{\mathbf{x}}(t), t). \end{aligned}$$

Since clearly $\tilde{\mathbf{x}}(t^{n+1}) = \check{\mathbf{x}}(t^{n+1}) = \mathbf{x}$, we have the claimed characteristic curve. Now

$$\begin{aligned} |\mathbf{u}(\mathbf{x},t) - \tilde{\mathbf{u}}(\mathbf{x},t)| &= \left|\mathbf{u}(\mathbf{x},t) - \mathbf{u}\left(\mathbf{x} - s\left(\frac{t^{n+1}-t}{\Delta t^n}\right)\boldsymbol{\nu}, t\right) + \frac{s}{\Delta t^n}\boldsymbol{\nu}\right| \\ &\leq \|\nabla \mathbf{u}\|_\infty |s| + \frac{|s|}{\Delta t^n} \leq C\frac{|s|}{\Delta t^n}, \end{aligned} \tag{3.68}$$

and by the uniform Lipschitz continuity of $\nabla \cdot \mathbf{u}$ in (3.4),

$$\begin{aligned} |\nabla \cdot \mathbf{u}(\mathbf{x},t) - \nabla \cdot \tilde{\mathbf{u}}(\mathbf{x},t)| &= \left|\nabla \cdot \mathbf{u}(\mathbf{x},t) - \nabla \cdot \mathbf{u}\left(\mathbf{x} - s\left(\frac{t^{n+1}-t}{\Delta t^n}\right)\boldsymbol{\nu}, t\right)\right| \\ &\leq L|s|\left|\frac{t^{n+1}-t}{\Delta t^n}\right| \leq L|s|. \end{aligned} \tag{3.69}$$

Combining (3.68) and (3.69) gives (3.67). □

**Lemma 3.5.5.** *Let $\tilde{E}^n(\alpha^*, s)$ be the trace-back polygonal approximation of $\check{E}^n$ with velocity field $\tilde{\mathbf{u}}$ defined in Lemma* 3.5.4 (*Figure* 3.3*, right*)*, and $V_{E^n}(\alpha^*, s) := |\tilde{E}^n(\alpha^*, s)|_\phi$ be its pore volume. Assume that no self-intersected polygons are created during the adjustment. If*

$$|\tilde{\mathbf{x}}_{\mathrm{l}}^n - \tilde{\mathbf{x}}_{\mathrm{r}}^n| \geq \lambda_* h \tag{3.70}$$

*for some constant $\lambda_* > 0$, then there exists some $s^*$ such that*

$$V_{E^n}(\alpha^*, s^*) = |\check{E}^n|_\phi, \tag{3.71}$$

*where $|s^*| \leq Ch\Delta t^n$ for some constant $C > 0$ independent of $n$, $h$, and $\Delta t^n$.*

*Proof.* For any $s$ in a neighborhood of zero, consider the difference

$$\begin{aligned} V_{E^n}(\alpha^*, s) - |\check{E}^n|_\phi = (V_{E^n}(\alpha^*, s) - V_{E^n}(\alpha^*, 0)) + (V_{E^n}(\alpha^*, 0) - V_{E^n}(0,0)) \\ + (V_{E^n}(0,0) - |\check{E}^n|_\phi). \end{aligned} \tag{3.72}$$

For the first term on the right hand side, since no self-intersected polygons are created during the adjustment, $\tilde{E}^n(\alpha^*, s)$ is monotone in $s$, so by (3.70),

$$|V_{E^n}(\alpha^*, s) - V_{E^n}(\alpha^*, 0)| \geq \frac{1}{2}\phi_* |\tilde{\mathbf{x}}_{\mathrm{l}}^n - \tilde{\mathbf{x}}_{\mathrm{r}}^n|\,|s| \geq \frac{1}{2}\phi_*\lambda_* h|s|. \tag{3.73}$$

For the second term on the right hand side of (3.72), notice that $\tilde{E}^n(\alpha,0) = \tilde{E}^n(\alpha) \subset \tilde{R}^n(\alpha)$ and the diameter of $\tilde{E}^n(\alpha)$ is $h_{\tilde{E}^n(\alpha)} \le \|\nabla \check{\mathbf{x}}\|_\infty h$, so by (3.52) and Lemma 3.5.2, we have

$$\begin{aligned} |V_{E^n}(\alpha^*,0) - V_{E^n}(0,0)| &\le |\tilde{E}^n(\alpha^*) \setminus \tilde{E}^n(0)| + |\tilde{E}^n(0) \setminus \tilde{E}^n(\alpha^*)| \\ &\le 2\big(h_{\tilde{E}^n(\alpha^*)} + h_{\tilde{E}^n(0)}\big)\|\mathbf{u}\|_\infty |\alpha^*| \Delta t^n \le C' h^2 \Delta t^n. \end{aligned} \tag{3.74}$$

For the third term on the right hand side of (3.72), since $\tilde{E}(0)$ is an octagon, by (3.61), we have

$$\big|V_{E^n}(0,0) - |\check{E}^n|_\phi\big| = \big||\tilde{E}(0)|_\phi - |\check{E}^n|_\phi\big| \le C' h^2 \Delta t^n. \tag{3.75}$$

Combining (3.72), (3.73), (3.74), and (3.75) gives

$$V_{E^n}(\alpha^*, s) - |\check{E}^n|_\phi \le \frac{1}{2}\phi_* \lambda_* h s + 2C' h^2 \Delta t^n < 0 \text{ when } s < -\frac{4C'}{\phi_* \lambda_*} h \Delta t^n, \tag{3.76}$$

$$V_{E^n}(\alpha^*, s) - |\check{E}^n|_\phi \ge \frac{1}{2}\phi_* \lambda_* h s - 2C' h^2 \Delta t^n > 0 \text{ when } s > \frac{4C'}{\phi_* \lambda_*} h \Delta t^n. \tag{3.77}$$

By the continuity of $V_{E^n}(\alpha^*, s) - |\check{E}^n|_\phi$, inequalities (3.76) and (3.77) imply that there exists some $s^*$, where $|s^*| \le Ch\Delta t^n$, such that equation (3.71) holds. □

*Remark* 3.5.5. If self-intersected polygons are created during the adjustment, one should reduce the distance $|s^*|$ in (3.71) by tracing and adjusting more points on an edge of a grid element. The assumption (3.70) implies that, again, the trace-back procedure should only be performed away from injection wells so that the length of segment $\tilde{\mathbf{x}}_{\mathrm{l}}^n \tilde{\mathbf{x}}_{\mathrm{r}}^n$ is non-degenerate.

Finally, combining Lemmas 3.5.1, 3.5.2, 3.5.4, and 3.5.5, we construct a perturbed velocity field $\tilde{\mathbf{u}}$ locally for all trace-back points, and they all have the $L^\infty$-error $\mathcal{O}(h)$ for $\mathbf{u}$ and $\nabla \cdot \mathbf{u}$. Then we can extend $\tilde{\mathbf{u}}$ to the entire domain $\Omega \times J_T$ by interpolating the local definitions of $\tilde{\mathbf{u}}$, and we keep the same bound for the error. In addition, due to the error $(\Delta t)^r$ of the approximately characteristic tracing in Remark 3.5.2, we obtain (3.6) and Assumption 3.0.1 holds.

## 3.6 Summary

The main result of this chapter is the proof of convergence of the fully conservative, volume corrected characteristics-mixed method for advection-diffusion equations without diffusion. Usually, we take the initial approximation $c_h^0 = P_h c^0$, which leads to an initial error $\|c_h^0 - c^0\|_1 = \mathcal{O}(h)$. The overall error is $\mathcal{O}(h/\sqrt{\Delta t} + h + (\Delta t)^r)$, where $r$ is related to the accuracy of the characteristic tracing itself (see Remark 3.5.2). In practice, we usually take the ratio $\Delta t/h$ to be a constant so the trace-back elements do not degenerate and self-intersect. Then the convergence rate of the method given by Theorem 3.4.1 is $\mathcal{O}(\sqrt{h})$. This rate is the same as Godunov's method, but we avoid the CFL constraint which puts an upper bound on the ratio $\Delta t/h$. Therefore, large time steps $\Delta t$ can be taken. However, as long as we do not introduce self-intersected trace-back regions, we can use much larger time steps. The optimal choice is $\Delta t = Ch^{2/(2r+1)}$, i.e., $\Delta t = Ch^{2/3}$ if $r = 1$, for a convergence rate $\mathcal{O}(h^{2/3})$. This is a better convergence rate than Godunov's method achieves.

The major difficulty of the proof is to verify the existence and error estimate of the locally conservative perturbed velocity field $\tilde{\mathbf{u}}$ in Assumption 3.0.1. Under some additional assumptions, our results guarantee that the volume correction step only produces a sufficiently small perturbation, and therefore maintains the convergence of the method. Actually, in practice, we do not calculate $\tilde{\mathbf{u}}$ or verify Assumptions 3.5.1–3.5.4. We just need to verify in the code that $\alpha^*$ and $s^*$ exist, which satisfy Lemmas 3.5.2 and 3.5.5, respectively, i.e., $\alpha^*$ and $s^*$ are not too large ($|\alpha^*| \leq Ch$ and $|s^*| \leq Ch\Delta t$).

# Chapter 4

# The Implementation of VCCMM

We cannot achieve locally conservative tracer simulation unless the velocity is also solved using a locally conservative method. Thus the flow problem must be solved in a locally conservative manner, such as a mixed finite element method [13, 45, 47] or a discontinuous Galerkin method [15, 26]. Then we will employ the locally mass and volume conserving characteristics method with the modification proposed by Arbogast and Huang [4] to solve for the tracer concentration $c$.

## 4.1 Flow Approximation

Generally the flow velocity $\mathbf{u}$ is a potential flow, given by combining the flow problem (2.1) and a constitutive equation, the simplest of which might be the Darcy's law

$$\mathbf{u} = -\mathbf{K}\nabla p, \tag{4.1}$$

where $p = p(\mathbf{x}, t)$ is the flow pressure and $\mathbf{K} = \mathbf{K}(\mathbf{x})$, in porous media applications, is the tensor of medium permeability divided by the fluid viscosity, which is assumed to be symmetric and uniformly positive definite.

Since the vector field $\mathbf{u}$ is the primary variable of interest, as it is used during the characteristic trace-back procedure, a locally conservative mixed finite element method is a good choice to approximate the velocity $\mathbf{u}$ and the pressure $p$ simultaneously to give approximating results for both variables.

For $(p, \mathbf{u}) = (p, \mathbf{u})(\cdot, t) \in W \times \mathbf{V}$, where $W$ and $\mathbf{V}$ are some scalar and vector function spaces on $\Omega$ with $\mathbf{u}(\cdot, t) \cdot \boldsymbol{\nu} = 0$ on $\partial\Omega$, respectively, the variational

form of system (2.1) and (4.1) is

$$(\nabla \cdot \mathbf{u}, w) = (q, w), \qquad w \in W, \tag{4.2}$$

$$(\mathbf{K}^{-1}\mathbf{u}, \mathbf{v}) = (p, \nabla \cdot \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}. \tag{4.3}$$

To discrete this system, let $W_h \subset W$ and $\mathbf{V}_h \subset \mathbf{V}$ be finite element spaces with basis functions $\{w_i\}_{1\le i\le N_W}$ and $\{\mathbf{v}_i\}_{1\le i\le N_V}$, respectively. For numerical solutions $(p_h, \mathbf{u}_h) = (p_h, \mathbf{u}_h)(\cdot, t) \in W_h \times \mathbf{V}_h$, we have the semi-discrete mixed finite element approximation of system (4.2)–(4.3)

$$(\nabla \cdot \mathbf{u}_h, w_h) = (q, w_h), \qquad w_h \in W_h, \tag{4.4}$$

$$(\mathbf{K}^{-1}\mathbf{u}_h, \mathbf{v}_h) = (p_h, \nabla \cdot \mathbf{v}_h), \quad \mathbf{v}_h \in \mathbf{V}_h. \tag{4.5}$$

Let $\mathbf{p}_h = \mathbf{p}_h(t) \in \mathbb{R}^{N_W}$ be the coefficient vector of $p_h$ represented as the linear combination of basis $\{w_i\}_{1\le i\le N_W}$, and $\vec{\mathbf{u}}_h = \vec{\mathbf{u}}_h(t) \in \mathbb{R}^{N_V}$ be the coefficient vector of $\mathbf{u}_h^k$ represented as the linear combination of basis $\{\mathbf{v}_i\}_{1\le i\le N_V}$. Then the vector form of system (4.4)–(4.5) is

$$\mathbf{B}^T\vec{\mathbf{u}}_h = \mathbf{q}, \tag{4.6}$$

$$\mathbf{A}\vec{\mathbf{u}}_h = \mathbf{B}\mathbf{p}_h, \tag{4.7}$$

or in a block matrix form

$$\begin{pmatrix} \mathbf{A} & -\mathbf{B} \\ -\mathbf{B}^T & \mathbf{0} \end{pmatrix} \begin{pmatrix} \vec{\mathbf{u}}_h \\ \mathbf{p}_h \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\mathbf{q} \end{pmatrix}, \tag{4.8}$$

where matrices $\mathbf{A} = (A_{i,j}) \in \mathbb{R}^{N_V \times N_V}$, $\mathbf{B} = (B_{i,j}) \in \mathbb{R}^{N_V \times N_W}$ and vector $\mathbf{q} = (q_i(t)) \in \mathbb{R}^{N_W}$ are defined as

$$\begin{aligned} A_{i,j} &:= (\mathbf{K}^{-1}\mathbf{v}_i, \mathbf{v}_j), \\ B_{i,j} &:= (w_j, \nabla \cdot \mathbf{v}_i), \\ q_i &:= (q, w_i). \end{aligned}$$

Eliminating $\vec{\mathbf{u}}_h$ in system (4.6)–(4.7) gives the symmetric linear system for $\mathbf{p}_h$

$$\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{p}_h = \mathbf{q}. \tag{4.9}$$

The solution to (4.9) is unique up to an additive vector in the kernel $\ker(\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}) = \text{span}\{\mathbf{e}\}$, where $\mathbf{e} = (1, 1, \cdots, 1)^T \in \mathbb{R}^{N_W}$. The linear system of algebraic equations (4.8) arising from the mixed finite element approximation are of *saddle type*, i.e., the system matrix has both positive and negative eigenvalues. Thus the solution of the system needs special care [18].

Now we will focus on the study of mixed finite element approximations for two dimensional rectangular meshes. Let $\Omega = (a_1, b_1) \times (a_2, b_2)$ and each interval $(a_i, b_i)$ be partitioned into $a_i = x_i^0 < x_i^1 < \cdots < x_i^{N_i} = b_i$, then the rectangular mesh $\mathcal{T}_h = \{E_{i,j} : 1 \le i \le N_1, 1 \le j \le N_2\}$, where the element $E_{i,j}$ is defined by $E_{i,j} := (x_1^{i-1}, x_1^i) \times (x_2^{j-1}, x_2^j)$.

### 4.1.1 $\mathbf{RT}_0$ approximation

**The $\mathbf{RT}_0$ space.** The lowest order Raviart-Thomas mixed finite element space ($\text{RT}_0$) [45] is defined by

$$\begin{aligned} W_h &:= \{w : w|_E \in Q_{0,0}(E), E \in \mathcal{T}_h\}, \\ \mathbf{V}_h &:= \{\mathbf{v} : \mathbf{v}|_E \in Q_{1,0}(E) \times Q_{0,1}(E), E \in \mathcal{T}_h; \mathbf{v} \cdot \boldsymbol{\nu} = 0 \text{ on } \partial\Omega \\ &\qquad \text{and } \mathbf{v} \cdot \boldsymbol{\nu} \text{ is continuous across edges of elements.}\}, \end{aligned}$$

where $Q_{k_1,k_2}(E)$ is the space of polynomials $p = p(x_1, x_2)$ on $E$ with degree up to $k_1$ in $x_1$ and $k_2$ in $x_2$. The degrees of freedom of $w \in W_h$ are determined by $\{w|_E : E \in \mathcal{T}_h\}$, and so

$$\dim(W_h) = N_W = N_1 N_2.$$

The degrees of freedom of $\mathbf{v} \in \mathbf{V}_h$ are determined by $\{\mathbf{v} \cdot \boldsymbol{\nu}_e|_e : e = (\partial E \cap \partial F) \setminus \partial\Omega, E, F \in \mathcal{T}_h\}$ (Fig. 4.1), where $\boldsymbol{\nu}_e$ is any unit vector orthogonal to edge $e$, and

$$\begin{aligned}\dim(\mathbf{V}_h) = N_V &= N_1(N_2 - 1) + (N_1 - 1)N_2 \\ &= 2N_1N_2 - (N_1 + N_2).\end{aligned}$$

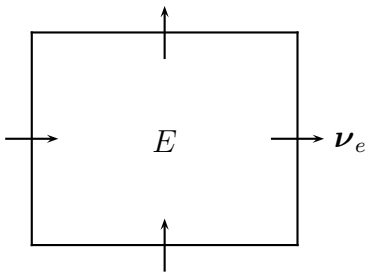


Figure 4.1: The degrees of freedom of $\mathbf{V}_h$ ($\mathrm{RT}_0$)

**Error estimate.** The $\mathrm{RT}_0$ mixed finite element approximation is first order accurate in $h$ for $p$ and $\mathbf{u}$ [18], i.e.,

$$\|p - p_h\|_{L^2} + \|\mathbf{u} - \mathbf{u}_h\|_{L^2} \leq C(\|p\|_{H^1} + \|\mathbf{u}\|_{H^1})h,$$

where $C > 0$ is a constant independent of $h$. By (4.4), there is no projection error for the divergence, i.e.,

$$\|P_h \nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_{L^2} = 0,$$

where $P_h$ is the $L^2$-projection operator defined in (3.10).

**Solvers.** Due to the simple structure of the $\mathrm{RT}_0$ space, system (4.6)–(4.7) can be approximated as a system generated by a *cell-centered* (or *block-centered*) finite difference scheme [48]. Thus $\mathbf{A}$ is approximated so as to be diagonal, and we easily obtain $\mathbf{A}^{-1}$. It is easy to form the operators $\mathbf{p} \mapsto \mathbf{u} = \mathbf{A}^{-1}\mathbf{B}\mathbf{p}$ and $\mathbf{u} \mapsto \mathbf{q} = \mathbf{B}^T\mathbf{u}$, so we can form the operator $\mathbf{p} \mapsto \mathbf{q} = \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\mathbf{p}$ and solve (4.9) by the method of Preconditioned Conjugate Gradients (PCG) [19]. Alternatively, we can form the

matrix $\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}$ and solve (4.9) by LAPACK routines [1] of direct linear system solvers. However, the $\mathrm{RT}_0$ approximation computes a discontinuous velocity field $\mathbf{u}_h \in V_h$.

### 4.1.2 $\mathrm{AW}_0$ approximation

**The $\mathrm{AW}_0$ space.** We study the improvement that may arise from using a fully conservative, continuous velocity field. Arbogast and Wheeler [6] introduced a family of rectangular mixed finite element spaces with a fully continuous flux, where the lowest order of such spaces ($\mathrm{AW}_0$) is defined by

$$W_h := \{w : w|_E \in Q_{0,0}(E), E \in \mathcal{T}_h\},$$
$$\mathbf{V}_h := \{\mathbf{v} \in C(\Omega)^2 : \mathbf{v}|_E \in Q_{1,2}(E) \times Q_{2,1}(E), E \in \mathcal{T}_h; \mathbf{v}\cdot\boldsymbol{\nu} = 0 \text{ on } \partial\Omega\}.$$

By adding more degrees of freedom to space $V_h$, the velocity field is fully continuous, but still allows control of the normal fluxes across the edges of elements. The degrees of freedom of $w \in W_h$ are determined by $\{w|_E : E \in \mathcal{T}_h\}$, and so

$$\dim(W_h) = N_W = N_1 N_2.$$

The degrees of freedom of $\mathbf{v} = (v_1, v_2)^T \in \mathbf{V}_h$ on each element $E_{i,j} \in \mathcal{T}_h$ are determined by the values of $\mathbf{v}$ at corner points of $E_{i,j}$ and midpoints of edges of $E_{i,j}$ (Fig. 4.2), i.e.,

$$\begin{gathered}
v_1(\mathbf{x}_{i-1,j-1}), \quad v_1(\mathbf{x}_{i-1,j-1/2}), \quad v_1(\mathbf{x}_{i-1,j}), \\
v_1(\mathbf{x}_{i,j-1}), \quad v_1(\mathbf{x}_{i,j-1/2}), \quad v_1(\mathbf{x}_{i,j}), \\
v_2(\mathbf{x}_{i-1,j-1}), \quad v_2(\mathbf{x}_{i-1/2,j-1}), \quad v_2(\mathbf{x}_{i,j-1}), \\
v_2(\mathbf{x}_{i-1,j}), \quad v_2(\mathbf{x}_{i-1/2,j}), \quad v_2(\mathbf{x}_{i,j}),
\end{gathered}$$

and $(\mathbf{v}\cdot\boldsymbol{\nu})(\mathbf{x}) = 0$ if $\mathbf{x} \in \partial\Omega$. So we have

$$\begin{aligned}
\dim(\mathbf{V}_h) = N_V &= (N_1 - 1)(2N_2 + 1) + (N_2 - 1)(2N_1 + 1) \\
&= 4N_1N_2 - 2(N_1 + N_2 + 1).
\end{aligned}$$

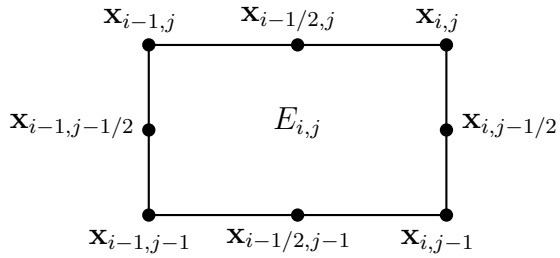


Figure 4.2: The degrees of freedom of $\mathbf{V}_h$ ($\mathrm{AW}_0$)

**Error estimate.** The $\mathrm{AW}_0$ mixed finite element approximation is first order accurate in $h$ for $p$, $\mathbf{u}$ and projection error of $\nabla \cdot \mathbf{u}$ [6], i.e.,

$$\|p - p_h\|_{L^2} + \|\mathbf{u} - \mathbf{u}_h\|_{L^2} + \|P_h \nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_{L^2} \leq C(\|p\|_{H^1} + \|\mathbf{u}\|_{H^1})h,$$

where $C > 0$ is a constant independent of $h$.

**Solvers.** For the $\mathrm{AW}_0$ space, the matrix $\mathbf{A}$ cannot be approximated by a diagonal matrix. Thus we must do more work to obtain $\mathbf{A}^{-1}$, which is a full matrix. Use of a direct solver is not reasonable. We propose using either PCG or Uzawa [14].

For PCG, we can form the operator in (4.9), $\mathbf{p} \mapsto (\mathbf{B}^T \mathbf{A}^{-1} \mathbf{B})\mathbf{p}$, in two stages: $\mathbf{Au} = \mathbf{Bp}$ solved using LAPACK for $\mathbf{u}$ and then computing $\mathbf{B}^T \mathbf{u}$.

Alternatively, for a given initial approximation $\mathbf{p}_{h,0} \in \mathbb{R}^{N_W}$ of $\mathbf{p}_h$, we have the Uzawa iteration sequence $(\mathbf{p}_{h,n}, \vec{\mathbf{u}}_{h,n})_{n \geq 0}$ given by

$$\vec{\mathbf{u}}_{h,n} = \mathbf{A}^{-1} \mathbf{B} \mathbf{p}_{h,n}, \tag{4.10}$$

$$\mathbf{p}_{h,n+1} = \mathbf{p}_{h,n} - \tau_U (\mathbf{B}^T \vec{\mathbf{u}}_{h,n} - \mathbf{q}), \tag{4.11}$$

where $\tau_U \in \mathbb{R}$ is the Uzawa parameter. Then

$$\lim_{n \to +\infty} (\mathbf{p}_{h,n}, \vec{\mathbf{u}}_{h,n}) = (\mathbf{p}_h, \vec{\mathbf{u}}_h)$$

for any $\mathbf{p}_{h,0} \in \mathbb{R}^{N_W}$ and a certain range of $\tau_U$.

## 4.2 Transport Approximation

### 4.2.1 Computation of characteristic trace-backs

To compute the characteristic trace-backs, we need to solve the ordinary differential equation for characteristics given by (2.5) with initial condition (2.6). In general, we cannot solve this equation analytically unless the velocity field $\mathbf{u}$ is particularly simple. Define the numerical interstitial velocity

$$\mathbf{v}_h = \frac{\mathbf{u}_h}{\phi_h}, \tag{4.12}$$

where $\mathbf{u}_h$ is the solution to the mixed finite element approximation (4.4)–(4.5) and $\phi_h$ is an approximation of $\phi \in C(\Omega)$.

#### 4.2.1.1 $\mathbf{RT}_0$ velocity field

The numerical interstitial velocity $\mathbf{v}_h$ is defined by (4.12) with $\phi_h = P_h\phi \in W_h(\Omega)$. Then $\mathbf{v}_h|_E \in Q_{1,0}(E) \times Q_{0,1}(E)$ for any $E \in \mathcal{T}_h$. Therefore, for each element $E$, the equation for characteristics (2.5) will be reduced to two uncoupled first order linear ordinary differential equations. That is, characteristic trace-back $\check{\mathbf{x}} = (\check{x}_1, \check{x}_2)^T \in E$ satisfies

$$\check{x}_i'(t) = a_i^E(t)\check{x}_i(t) + b_i^E(t), \quad i = 1, 2,$$

which can be solved analytically. This can give us accurate characteristic trace-backs. However, the velocity field $\mathbf{v}_h$ is discontinuous across the edges of elements, so great care must be taken when we trace back a grid point or a point is traced to an edge or a corner of an element [38].

#### 4.2.1.2 $\mathbf{AW}_0$ velocity field

For the fully continuous velocity field $\mathrm{AW}_0$ [6], define the numerical interstitial velocity $\mathbf{v}_h$ by (4.12) with $\phi_h = \phi \in C(\Omega)$. Then $\mathbf{v}_h \in C(\Omega)^2$ and many numerical methods can be employed to solve equation (2.5). The most straightforward way is to use the multi-step forward Euler method to trace points back

along the direction of the velocity. To do so, for each time step $J^n$, one makes a subgrid of interval $J^n$, $t^n = t^{n,0} < t^{n,1} < \cdots < t^{n,N_n} = t^{n+1}$ with time substep $\Delta t^{n,k} := t^{n,k+1} - t^{n,k}$, and computes the following sequence backward in time

$$\check{\mathbf{x}}(t^{n,k}) = \check{\mathbf{x}}(t^{n,k+1}) - \Delta t^{n,k} \mathbf{v}_h(\check{\mathbf{x}}(t^{n,k+1}), t^{n,k+1})$$

with $\check{\mathbf{x}}(t^{n,N}) = \check{\mathbf{x}}(t^{n+1}) = \mathbf{x}$ given.

### 4.2.2 Adjustment of trace-back points

As described in the Volume Correction Algorithm in Section 2.2, we adjust trace-back points in time. During each of the trace-forward injection well adjustment (Step 1), ring adjustment (Step 2) and individual element adjustment (Step 3), bisection with a cut factor in time will be used.

#### 4.2.2.1 Algorithm of trace-back points adjustment

Let $V_0$ be a target volume that we need to obtain, $\tilde{V}$ be the adjusted volume, $\varepsilon_{\text{tol}}$ be the relative tolerance error of $\tilde{V}$, and $\lambda_{\text{cut}}$ be the cut factor of time ($0 < \lambda_{\text{cut}} < 1$).

**Steps 1, 2.** The trace-forward boundary of an injection well or trace-back boundary of a ring away from an injection well is adjusted in time according to the relative error of adjusted volume $\tilde{V}$ until the volume conservation is achieved. A brief algorithm is depicted in Figure 4.3.

**Step 3.** After a trace-back ring $\tilde{R}^n$ is adjusted, each individual element $\tilde{E}^n \subset \tilde{R}^n$ is adjusted by moving the midpoint of an edge in the traverse direction of the flow. To avoid introducing systematic bias of volume errors into the adjusted elements, the target volume of $\tilde{E}^n$ to be adjusted should be $V_0(\tilde{E}^n) = |E|(1 + \varepsilon_{\tilde{R}^n})$ instead of $|E|$, where $\varepsilon_{\tilde{R}^n}$ is the relative (rounding) error of the volume of $\tilde{R}^n$ after adjustment in Step 2. The definition of

$V_0(\tilde{E}^n)$ keeps the tessellation of volumes in $\tilde{R}^n$ up to rounding errors, i.e.,

$$\sum_{E \subset R} V_0(\tilde{E}^n) = \sum_{E \subset R} |E|(1 + \varepsilon_{\tilde{R}^n}) = |R|(1 + \varepsilon_{\tilde{R}^n}) = |\tilde{R}^n|.$$

$\varepsilon \leftarrow (\tilde{V} - V_0)/V_0$;
$\tau \leftarrow \lambda_{\text{cut}} \Delta t^n$;
**while** $|\varepsilon| > \varepsilon_{\text{tol}}$
    **if** $\varepsilon > 0$
        trace backward for time $\tau$;
    **else**
        trace forward for time $\tau$;
    **end if**
    update $\tilde{V}$;
    $\varepsilon \leftarrow (\tilde{V} - V_0)/V_0$;
    $\tau \leftarrow \lambda_{\text{cut}} \tau$;
**end while**

Figure 4.3: Bisection algorithm with a cut factor $\lambda_{\text{cut}}$ in time.

In practice, this is an efficient way to correct the trace-back volumes since we use a potential velocity field $\mathbf{u}$ which is given by the Darcy's law (4.1), so the exact trace-back region $\check{E}^n$ would not distort too much, which means the polygonal approximation $\tilde{E}^n$ should be close to $\check{E}^n$ even without the adjustment if $\Delta t$ and $h$ are chosen properly.

#### 4.2.2.2 Polyline structure

During the adjustment, we frequently compute the volumes of polygons, such as trace-back rings $\tilde{R}^n$ and elements $\tilde{E}^n$. Our code features a class of polyline structure (Figure 4.4), which uses a doubly linked list of vertices to represent a polygon. This structure gives us the connectivity of vertices, which is convenient to compute volumes of polygons. If the vertices $\{\mathbf{x}_i\}_{1 \le i \le n} \subset \mathbb{R}^2$ of a polygon

$G$ are given in clockwise or counterclockwise direction, then its volume can be computed by the formula

$$|G| = \frac{1}{2}\left|\sum_{i=1}^{n} \det(\mathbf{x}_i, \mathbf{x}_{i+1})\right|, \tag{4.13}$$

where $\mathbf{x}_{n+1} = \mathbf{x}_1$.

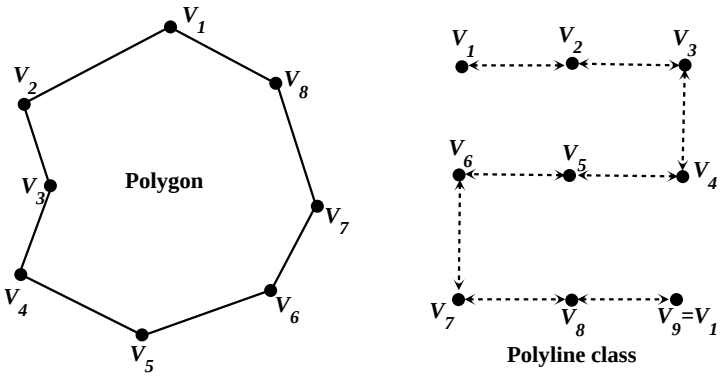


Figure 4.4: Polyline structure
Left: A polygon which approximates a trace-back region. Right: The polyline structure, which is a doubly linked list of vertices, represents the polygon.

### 4.2.3 Update of the numerical solution

For each time step $J^n$, after we correct the volume of each trace-back element $\tilde{E}^n$, the update of the numerical solution $c_h^{n+1}$ can be computed by the vector form of VCCMM scheme (2.22). To do so, we need to compute the matrix $\mathbf{A}_h^n$ and vector $\mathbf{b}_h^n$ defined by (2.23) and (2.24), respectively. The difficulty of the computation is to compute the volume $|\tilde{E}^n \cap F|$, where $F$ is an element intersecting $\tilde{E}^n$. Actually, the polygon $\tilde{E}^n \cap F$ can be calculated by the Sutherland-Hodgman clipping algorithm [34, pp. 124-127], where we consider $\tilde{E}^n$ as the subject polygon which is clipped by a rectangular clipping window $F$ (Figure 4.5).

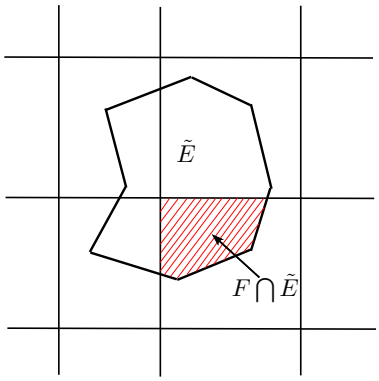


Figure 4.5: A trace-back element $\tilde{E}$ is clipped by a grid element $F$.

This clipping algorithm works by extending each edge of the clipping window $F$ in turn and selecting only intersection points and vertices from the subject polygon $\tilde{E}^n$ that are on the "visible" side, which is the side $F$ lies with respect to its extended edge. After $\tilde{E}^n$ is clipped by each side of $F$, the algorithm generates a sequence of vertices which are those of the clipped polygon $\tilde{E}^n \cap F$ given in clockwise or counterclockwise direction if the vertices of $\tilde{E}^n$ are given so. Then the volume of $\tilde{E}^n \cap F$ is computed by (4.13).

It is convenient to implement the Sutherland-Hodgman clipping algorithm with our class of polyline structure, since it allows us to advance through the edges of a polygon in turns, which is consistent with the feature of this clipping algorithm. It also provides the flexibility to applications of more general and complicated meshes.

# Chapter 5

# Computational Tests

In this chapter, we demonstrate some computational tests solved by VC-CMM with the implementation in Chapter 4, and compare the numerical results with those of the CMM and Godunov's method.

## 5.1 Rotating Pollutant Problem

To show the accumulation effect of projection errors, we test an example where numerical diffusion cannot be tolerated, and so we must avoid using small time steps if possible.

We consider the following system for the concentration $c$ of a pollutant

$$\begin{aligned} c_t + \nabla \cdot (c\mathbf{u}) &= 0 \qquad \text{in } \mathbb{R}^2 \times [0, 2\pi], \\ c(\mathbf{x}, 0) &= \chi_E(\mathbf{x}) \qquad \text{in } \mathbb{R}^2, \end{aligned}$$

where the velocity field $\mathbf{u}(\mathbf{x}) = (-x_2, x_1)$ is rotating counterclockwise around the origin $O = (0, 0)$, and $\chi_E$ is the characteristic function of $E \subset \mathbb{R}^2$ defined as $\chi_E(\mathbf{x}) = 1$ if $\mathbf{x} \in E$ and 0 if $\mathbf{x} \notin E$. We use the uniform square mesh on integers in $\mathbb{R}^2$, and let $E = (9, 10) \times (0, 1)$ be the initial polluted area. Since $\mathbf{u}$ is simply a rotation, the polluted area should be the same after time $T = 2\pi$. For $N$ time steps within time $T$, we plot the maximum concentrations of the pollutant in Figure 5.1 and concentration profiles at time $T$ in Figure 5.2, respectively. These pictures show that, as $N$ increases, the maximum concentration decreases and the polluted area expands rapidly due to an increase of numerical diffusion.

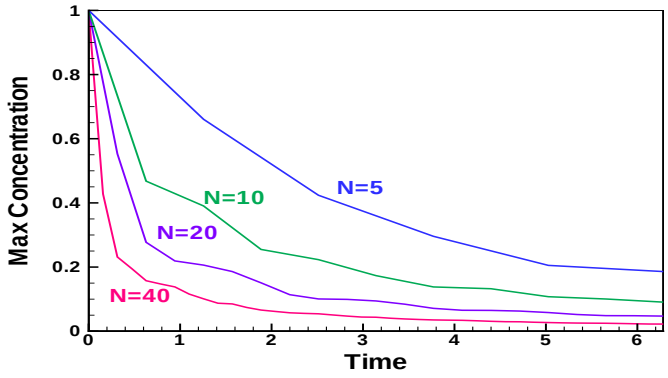


Figure 5.1: Maximum concentration of pollutant

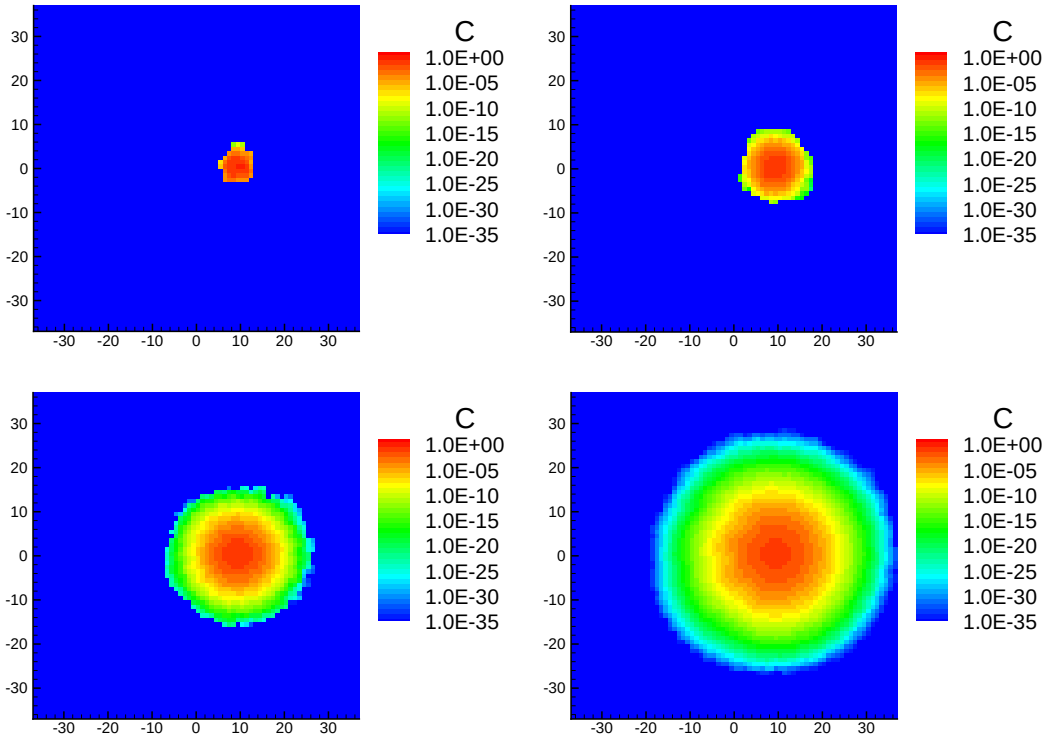


Figure 5.2: Pollutant concentration at time $T = 2\pi$. Shown are $N = 5$ (upper left), $N = 10$ (upper right), $N = 20$ (lower left), and $N = 40$ (lower right).

## 5.2 Convergence Tests of VCCMM

To verify the convergence rate in Theorem 3.4.1, we consider a quarter of a "five-spot" pattern of wells, which is a rectangular domain $\Omega = (0, 15) \times (0, 20)$ meters with a tracer injection well near the corner $(0, 0)$, a production well near the corner $(15, 20)$, and boundary condition (2.9). We impose a uniform $n \times n$ rectangular grid over $\Omega$ and a uniform time step $\Delta t$. It is initially clean: $c^0(\mathbf{x}) = 0$. The injector covers one cell near the corner $(0, 0)$ and has a constant rate of $q = 1.2$ m$^2$/minute, injecting an inert tracer with concentration $c_I = 1$. The cell comprising the producer near the opposite corner $(15, 20)$ has rate opposite that of the injector. We assume the fluid viscosity $\mu = 0.01$ poise is constant (i.e., the concentration of the tracer is too small to affect the viscosity of the fluid, which is water). For simplicity, we solve (2.7) with a constant porosity $\phi(\mathbf{x}) \equiv 1$ and a uniform isotropic permeability tensor $\mathbf{K}(\mathbf{x}) = k(\mathbf{x})\mathbf{I}$, where $\mathbf{I}$ is the identity tensor, and $k(\mathbf{x}) \equiv 10$ millidarcies.

To test the optimal convergence rate with Euler's method for solving characteristics (i.e., $r = 1$ in Theorem 3.4.1), let $\Delta t = Ch^{2/3}$ and compute the normalized discrete $L^\infty(J_T; L^1(\Omega))$-error

$$E_h := \frac{1}{|\Omega|} \max_{0 \le k \le N} \|c_h^k - c^k\|_1 \tag{5.1}$$

in Theorem 3.4.1. We approximate (2.8)–(2.10) using VCCMM for the simulation time $T = 1$ hour, and consider the "exact" solution $c$ computed by the higher order Godunov's method [10, 25] on a fine $256 \times 256$ grid using the restricted CFL time step $\Delta t_{\text{CFL},256} \approx 0.23$ second. Table 5.1 shows the error $E_{h_n}$ and the ratio $C_{h_n} := E_{h_n}/h_n^{2/3}$ on grids for 6 different sizes $n$. From the results, the sequence of the ratio $C_{h_n}$ shows an upper bound $C^*$ as $h_n$ decreases to zero, so indeed

$$E_{h_n} \le C^* h_n^{2/3},$$

which is consistent with Theorem 3.4.1, and indicates that VCCMM is convergent and has the optimal convergence rate of at least $\mathcal{O}(h^{2/3})$.

| $n$ | $h_n$ (m) | $\Delta t_n$ (sec) | $E_{h_n}$ | $C_{h_n}$ |
|---|---|---|---|---|
| 8 | 3.1250 | 115.35 | 0.46870 | 0.2193 |
| 16 | 1.5625 | 72.66 | 0.19137 | 0.1421 |
| 32 | 0.7813 | 45.78 | 0.07837 | 0.0924 |
| 64 | 0.3906 | 28.84 | 0.03507 | 0.0656 |
| 128 | 0.1953 | 18.17 | 0.01767 | 0.0525 |
| 256 | 0.0977 | 11.44 | 0.00882 | 0.0416 |

Table 5.1: Convergence test 1 for $\Delta t = Ch^{2/3}$. The sequence of $C_{h_n} \leq C^*$, so $E_{h_n} \leq C^* h^{2/3}$.

The next test indicates that $\mathcal{O}(h^{2/3})$ is exactly the optimal convergence rate of VCCMM when $\Delta t = Ch^{2/3}$. Consider a constant velocity field $\mathbf{u} \equiv (0.03, 0.04)$ m/sec and no source or sink (i.e., $q = 0$). Then the in-flow boundary $\Gamma_{\text{in}}$ is the union of the left and bottom edges of $\Omega$ (Figure 5.3). We impose the boundary and initial conditions

$$c(\mathbf{x}, t) = 1 \text{ on } \Gamma_{\text{in}} \times J_T \quad \text{and} \quad c^0(\mathbf{x}) = 0 \text{ in } \Omega,$$

where $T = 500$ seconds. Then the exact solution $c$ is

$$c(\mathbf{x}, t) = \begin{cases} 1 & \text{if } x_1 \leq 0.03\, t \text{ or } x_2 \leq 0.04\, t, \\ 0 & \text{otherwise,} \end{cases}$$

and the flow front is shown as the dashed line in Figure 5.3 with the corner point $\mathbf{x}_{\text{f}}(t) = t\mathbf{u}$. In paticular, at time $T$, $\mathbf{x}_{\text{f}}(t) = (15, 20)$, and the entire domain $\Omega$ is flooded.

Due to the simplicity of $\mathbf{u}$, there is no need for the polygonal approximation and volume adjustment procedures of VCCMM. Table 5.2 shows the error $E_{h_n}$ defined in (5.1) and the ratio $C_{h_n} := E_{h_n}/h_n^{2/3}$ with grids of 7 different sizes $n$. From the results, the sequence of the ratio $C_{h_n}$ is stable around 0.03 as $h_n$ decreases to zero, so the optimal convergence rate is apparently exactly $\mathcal{O}(h^{2/3})$ as expected from Theorem 3.4.1.

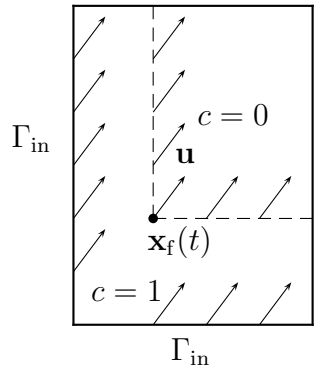


Figure 5.3: A domain flooded by the flow with a constant velocity $\mathbf{u}$

| $n$ | $h_n$ (m) | $\Delta t_n$ (sec) | $E_{h_n}$ | $C_{h_n}$ |
|---|---|---|---|---|
| 8 | 3.1250 | 166.67 | 0.07252 | 0.0339 |
| 16 | 1.5625 | 105.00 | 0.04508 | 0.0335 |
| 32 | 0.7813 | 66.14 | 0.02506 | 0.0295 |
| 64 | 0.3906 | 41.67 | 0.01639 | 0.0307 |
| 128 | 0.1953 | 26.25 | 0.00972 | 0.0289 |
| 256 | 0.0977 | 16.54 | 0.00670 | 0.0316 |
| 512 | 0.0488 | 10.42 | 0.00396 | 0.0297 |
| 1024 | 0.0244 | 6.56 | 0.00262 | 0.0311 |

Table 5.2: Convergence test 2 for $\Delta t = Ch^{2/3}$. The sequence of $C_{h_n} \approx C^* = 0.03$, so $E_h \approx C^* h^{2/3}$.

## 5.3 Comparison of $\mathbf{RT}_0$ and $\mathbf{AW}_0$ Flow Approximations

Now we use a heterogeneous permeability $k(\mathbf{x})$ depicted in Figure 5.4, that is geostatistically generated and has mean $m_k = 10$ md and dimensionless coefficient of variation $C_v = 2.5$, where

$$m_k = \frac{1}{|\Omega|}\int_\Omega k(\mathbf{x})\,d\mathbf{x}, \quad C_v = \frac{1}{m_k}\left(\frac{1}{|\Omega|}\int_\Omega (k(\mathbf{x}) - m_k)^2 d\mathbf{x}\right)^{1/2}.$$

The permeability $k(\mathbf{x})$ varies by about 4 orders of magnitude ($10^{-2}$ to $10^2$ md).

Figure 5.5 shows the divergences of velocity using $\text{RT}_0$ and $\text{AW}_0$ mixed finite element approximations on a $50 \times 50$ grid. At wells, both of them compute

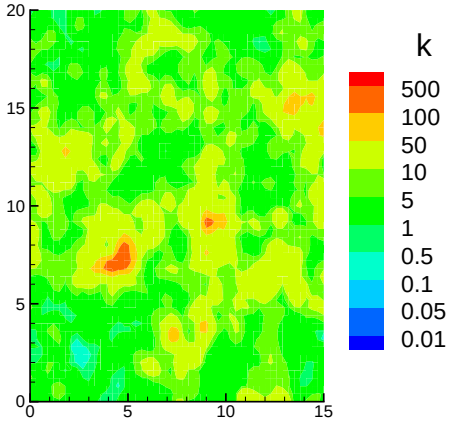


Figure 5.4: A heterogenous permeability field in millidarcies (md)

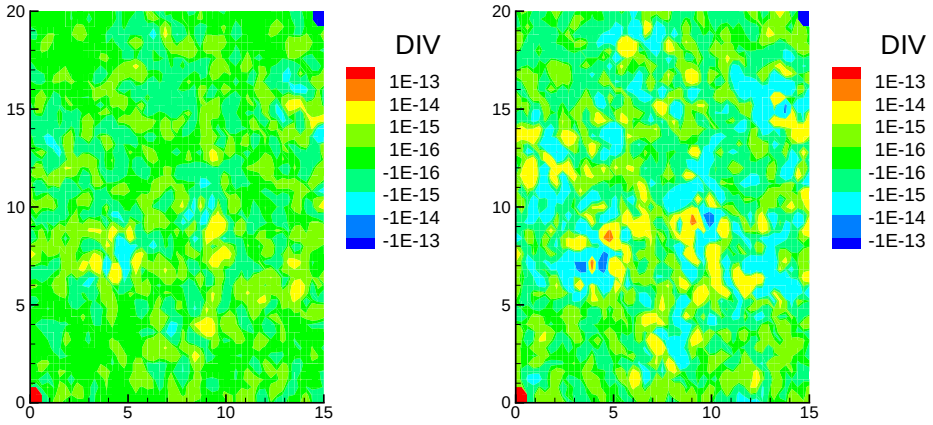


Figure 5.5: Divergence of velocity in $\text{sec}^{-1}$ (Grid: $50 \times 50$). The left shows $\text{RT}_0$, the right, $\text{AW}_0$.

a divergence of velocity $\nabla \cdot \mathbf{u} = 0.1667$ $\text{sec}^{-1}$ for the injector and $\nabla \cdot \mathbf{u} = -0.1667$ $\text{sec}^{-1}$ for the producer. This is consistent, since the well rate per unit pore volume

$$\frac{\text{well rate}}{\text{well volume}} = \frac{1.2 \text{ m}^2/\text{min}}{0.12 \text{ m}^2} = 10 \text{ min}^{-1} \approx 0.1667 \text{ sec}^{-1}.$$

Due to the local heterogeneity of the permeability, the divergence on the rest of the field varies by about 4 orders of magnitude ($10^{-16}$ to $10^{-13}$ $\text{sec}^{-1}$). The $\text{RT}_0$ approximation satisfies the conservation of bulk fluid (2.1) pointwise, but the $\text{AW}_0$

approximation only satisfies (2.1) on the average in each grid element. Thus, the divergence for $RT_0$ shows a better accuracy than that for $AW_0$. However, these errors are extremely small, so that they do not affect the quality of the transport approximations, as show in Figures 5.6.

Figure 5.6 shows the tracer concentration profiles at time $t = 100$ minutes with time step $\Delta t = 30$ seconds using $RT_0$ and $AW_0$ flow approximations. The concentration profiles computed with flows approximated by $RT_0$ and $AW_0$ show basically the same quality, but the latter one shows less numerical diffusion, in that the flooding front is sharper, perhaps, due to the higher order approximation of $\mathbf{u}$ using $AW_0$.

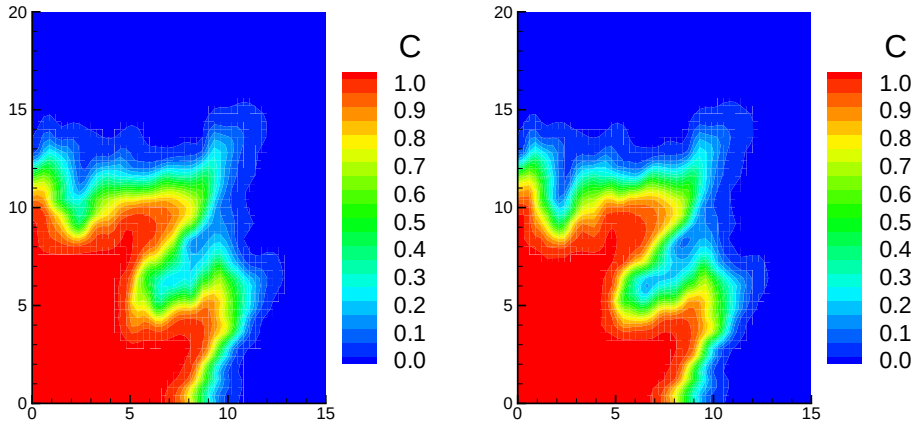


Figure 5.6: Tracer concentration at time $t = 100$ min (Grid: $50 \times 50$). The left shows $RT_0$, the right, $AW_0$.

## 5.4 Comparison of VCCMM with CMM and First Order Godunov's Method

We use the same parameters given in Section 5.3. Figure 5.7 shows the tracer concentration profiles at time $t = 100$ minutes computed with a $50 \times 50$ grid by unmodified CMM, first order Godunov's method (FOG), and VCCMM

with $RT_0$ and $AW_0$ flow approximations. We use the CFL restricted $\Delta t_{\text{CFL}} = 6$ seconds for the FOG and $\Delta t = 80$ seconds for other methods.

Due to the violation of the local volume conservation, the CMM exhibits both overshoots and undershoots, and introduces many nonphysical local minima and maxima into the solution. The FOG and VCCMM correct these problems and exhibit no undershoot or overshoot, which give solutions with monotone values of contours in the field. However, the VCCMM is less numerically diffuse than FOG due primarily to using much larger time steps ($\Delta t \approx 13.3 \Delta t_{\text{CFL}}$).

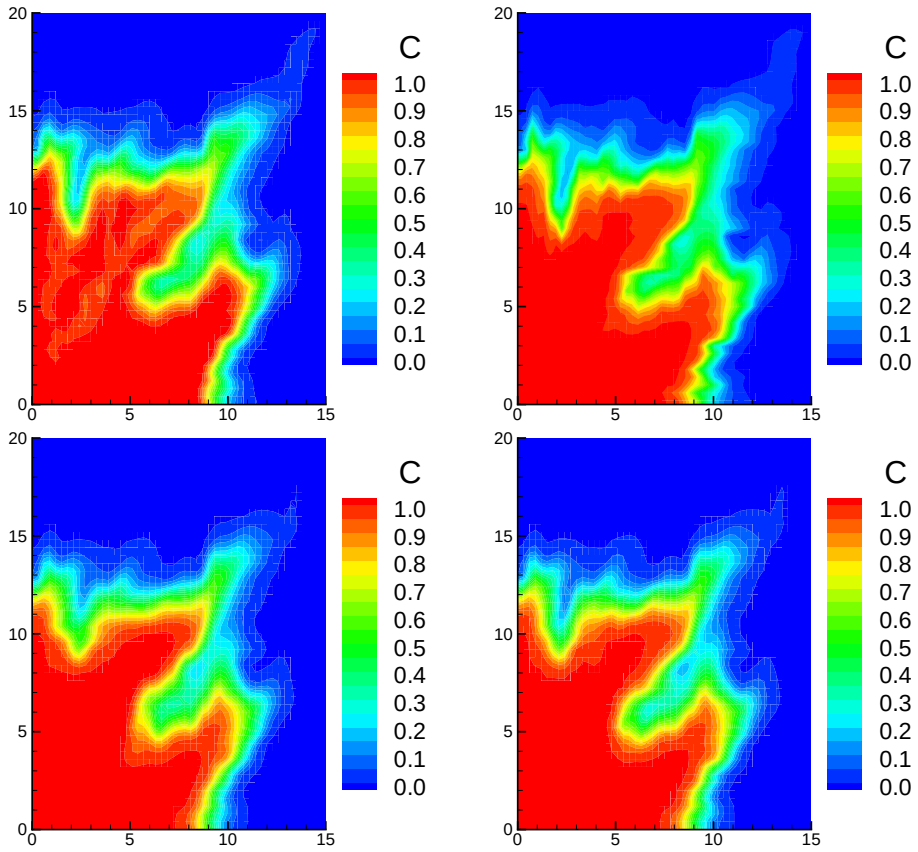


Figure 5.7: Tracer concentration at time $t = 100$ min (Grid: $50 \times 50$). Shown are CMM (upper left), FOG (upper right), VCCMM-$RT_0$ (lower left), and VCCMM-$AW_0$ (lower right).

Figure 5.8 shows the results by using a refined $100 \times 100$ grid. We use the restricted $\Delta t_{\text{CFL}} = 1.5$ seconds for FOG and $\Delta t = 30$ seconds for CMM and VCCMM. Again, the CMM exhibits both overshoots and undershoots. Each of FOG and VCCMM corrects these problems and predicts a similar front shape. However, the VCCMM-$\text{AW}_0$ is less numerically diffuse than FOG and VCCMM-$\text{RT}_0$ due to using much larger time steps ($\Delta t = 20\Delta t_{\text{CFL}}$) and higher order approximation of flows.

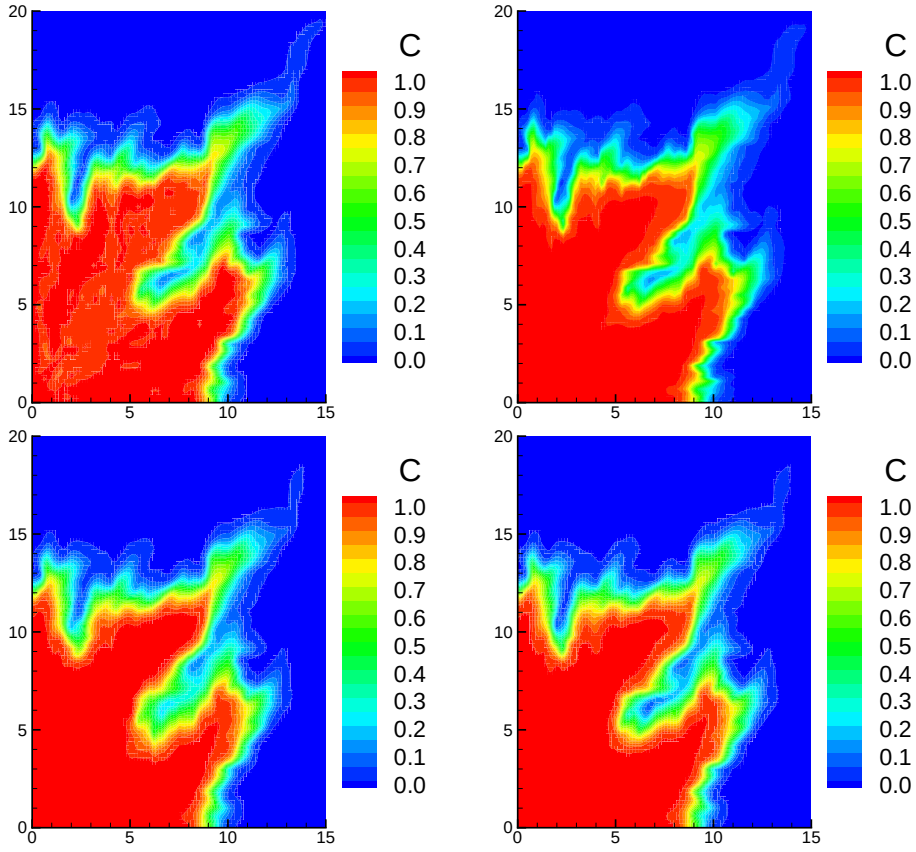


Figure 5.8: Tracer concentration at time $t = 100$ min (Grid: $100 \times 100$). Shown are CMM (upper left), FOG (upper right), VCCMM-$\text{RT}_0$ (lower left), and VCCMM-$\text{AW}_0$ (lower right).

## 5.5 Summary of Conputational Tests

The computational tests demonstrated in this chapter compared the results of the VCCMM with the CMM and Godunov's method, indicating that the VCCMM avoided non-physical oscillations in solutions, and we should use larger time steps in the computation, if possible, to reduce the accumulation of projection errors. Actually, large time steps ($13.3\Delta t_{\mathrm{CFL}}$–$20\Delta t_{\mathrm{CFL}}$) were taken in practice compared with the Godunov's method. The $\mathrm{RT}_0$ and $\mathrm{AW}_0$ approximations gave similar results, but the latter showed more accurate results by using higher order approximations of flows. In addition, the VCCMM computed an accurate solution compared with the higher order Godunov's method on a fine grid, and matched the optimal convergence rate in Theorem 3.4.1.

# Chapter 6

# An Application of VCCMM to a Nuclear Waste Disposal Simulation

In this chapter, we consider a simplified Far Field realistic model aimed at simulating the transport of radionuclides around a nuclear waste repository. The problem was defined originally by ANDRA [11] in the early 2000's used for safety assessments in nuclear waste management. It leads to a classical advection-diffusion-reaction type problem. Since we are demonstrating VCCMM, we make some modifications to the problem to better match the limitations of our demonstration code.

## 6.1 The Problem

From the mathematical point of view, the problem is modelled by an advection-diffusion-reaction equation with some boundary and initial conditions. However, the physical parameters in the equation are highly varying from one layer to another, and the source is highly concentrated in space and time. In addition, it is an extremely long time (i.e., millions of years) that both phenomena of advection and diffusion are active.

### 6.1.1 The computational domain and layers

The computation is restricted to a two-dimensional rectangular disposal site $\Omega = (0, 25000) \times (0, 695)$ in meters on the $xy$-plane with layers of dogger ($k_{\mathrm{h}} = 25.2288$ m/year), clay ($k_{\mathrm{h}} = 3.1536 \times 10^{-6}$ m/year), limestone ($k_{\mathrm{h}} = 6.3072$

m/year), and marl ($k_{\mathrm{h}} = 3.1536 \times 10^{-5}$ m/year). In this chapter, we use a constant layer hydraulic conductivity[1] $k_{\mathrm{h}}$ in meter/year, as depicted in Figure 6.1. Note that $k_{\mathrm{h}}$ varies about 7 orders of magnitude. A deep geological repository, denoted by $R$, is modelled by a rectangular region in the clay layer with dimensions $R = (18440, 21680) \times (244, 250)$ meters shown in red. The computation is carried over time $J_T$ with $T = 10^6$ years.

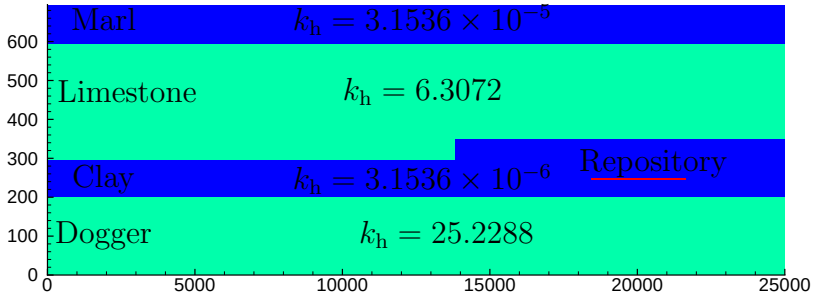


Figure 6.1: The computational domain of the disposal site showing four layers and the repository (shown in red).

### 6.1.2 The flow

It is assumed that all rock layers are saturated with water and that boundary conditions are stationary, so that the flow and pressure are independent of time. Darcy's law gives the velocity

$$\mathbf{u} = -k_{\mathrm{h}} \nabla \Psi$$

in terms of the hydrodynamic load $\Psi := p/(\rho|\mathbf{g}|) + y$, where we assume the density $\rho$ is a constant, and $\mathbf{u}$ satisfies the mass conservation (2.1) with $q = 0$. Also, we

[1] For a subsurface system, hydraulic conductivity can be expressed as $k_{\mathrm{h}} = k\rho|\mathbf{g}|/\mu$ [9, pp. 133], where $k$ is the medium permeability, $\rho$ is the fluid density, $\mathbf{g}$ is the gravitational acceleration, and $\mu$ is the fluid viscosity.

impose the following boundary conditions in meters

$$\begin{aligned}
\Psi &= 286 && \text{on } \{0\} \times (0, 200),\\
\Psi &= 200 && \text{on } \{0\} \times (295, 595),\\
\Psi &= 289 && \text{on } \{25000\} \times (0, 200),\\
\Psi &= 310 && \text{on } \{25000\} \times (350, 595),\\
\nabla\Psi \cdot \boldsymbol{\nu} &= 0 && \text{elsewhere.}
\end{aligned}$$

### 6.1.3 The governing equation for transport

At the initial time, the repository has a leak, and we consider the long-lived radioactive element iodine 129 that escapes from the repository cave into the water. The leak maintains a repository concentration $c^0 = 0.133$ mol/m$^3$. The concentration $c$ is given by the advection-diffusion-reaction equation

$$\phi(c_t + \lambda c) + \nabla \cdot (c\mathbf{u} - \mathbf{D}\nabla c) = 0 \qquad \Omega \times J_T, \tag{6.1}$$

where the effective porosity $\phi = 0.001$ in the clay layer and 0.1 elsewhere, the radioactive decay constant $\lambda = \log(2)/T_{\text{half}}$ with the half life time $T_{\text{half}} = 1.57 \times 10^7$ years, and the effective diffusion/dispersion tensor $\mathbf{D}$ depends on the Darcy velocity $\mathbf{u}$ as

$$\mathbf{D}(\mathbf{u}) = \phi d_{\text{mol}}\mathbf{I} + |\mathbf{u}|\,[d_{\text{long}}E(\mathbf{u}) + d_{\text{trans}}(I - E(\mathbf{u}))],$$

where $E(\mathbf{u}) = \mathbf{u}\mathbf{u}^T/|\mathbf{u}|^2$ and molecular diffusion, longitudinal and transverse dispersion coefficients, assumed constant in each layer, are given in Table 6.1 below. Also, we impose the boundary conditions for transport as

$$\begin{aligned}
\nabla c \cdot \boldsymbol{\nu} &= 0 && \text{on } \{0\} \times \{(0, 200) \cup (295, 595)\},\\
c &= 0 && \text{elsewhere.}
\end{aligned}$$

| | $d_{\text{mol}}$ ($\text{m}^2$/year) | $d_{\text{long}}$ (m) | $d_{\text{trans}}$ (m) |
|---|---|---|---|
| Dogger | $5 \times 10^{-4}$ | 50 | 1 |
| Clay | $9.48 \times 10^{-7}$ | 0 | 0 |
| Limestone | $5 \times 10^{-4}$ | 50 | 1 |
| Marl | $5 \times 10^{-4}$ | 0 | 0 |

Table 6.1: Diffusion/dispersion coefficients in the four layers.

## 6.2 Numerical Method

We use a non-uniform $108 \times 70$ rectangular mesh with the local refinement near the repository shown in Figure 6.2. We use an operator splitting technique to solve (6.1) by approximating the advection with Godunov's method or VCCMM, approximating the diffusion with the expanded mixed finite element method [17], and solving the reaction analytically by solving an ordinary differential equation.

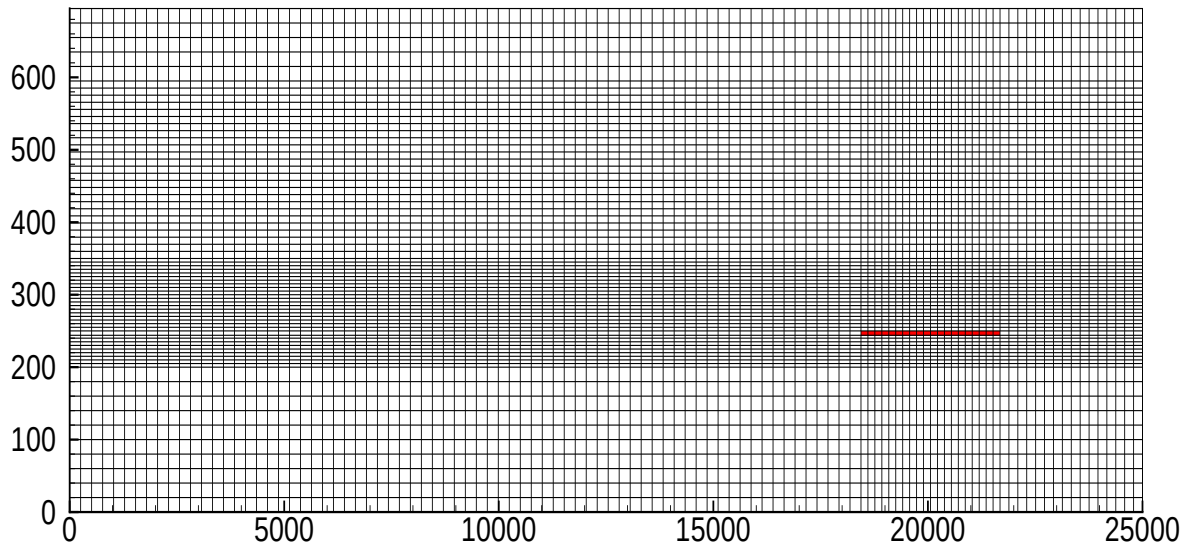


Figure 6.2: Non-uniform $108 \times 70$ rectangular mesh with local refinement near the repository.

## 6.3 Flow Approximation Results

Figure 6.3 shows the hydrodynamic load (top) and speed (bottom) with the $RT_0$ mixed finite element approximation. There is approximately a linear hydrodynamic load drop in the limestone layer since the conductivity is constant. The clay layer shows a nearly constant hydrodynamic load since the hydrodynamic load drop in this layer is small. For the speed, as we should expect, there is a relatively large speed in the limestone layer since, not only is this layer more permeable, but also there is a larger hydrodynamic load drop on the boundaries. Clay and marl layers have little speed due to low conductivities and no-flow boundary conditions.

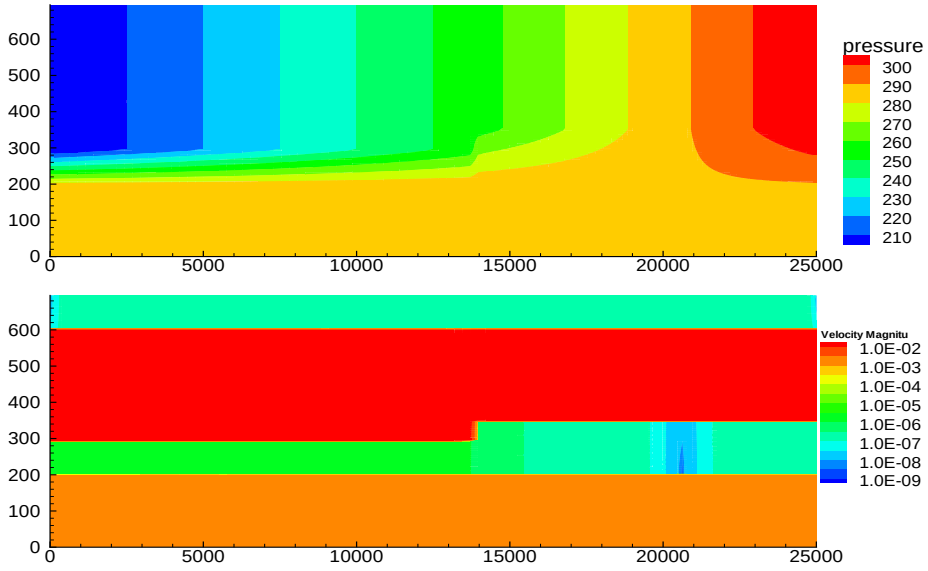


Figure 6.3: Flow approximation of the hydrodynamic load (top) and speed (bottom).

## 6.4 Transport Approximation Results

Due to the simple structure of the conductivity distribution in the domain, we are able to use a large time step $\Delta t = 2500$ years for VCCMM, and only

little work is needed for the trace-back adjustment. We take $\Delta t = 100$ years for Godunov's method because the CFL restricted $\Delta t_{\text{CFL}} \approx 102.52$ years.

Figure 6.4 shows the characteristic trace-back mesh for time $\Delta t = 2500$ years, where we treat the inflow boundary $\{25000\} \times (0, 695)$ as an injection well using the trace-forward technique. Due to the large time step, there are some self-intersected trace-back polygons created near the "sharp corner" of the interface between the clay and limestone layers. However, this degeneracy only results in a locally minor inaccuracy of the transport approximation as shown in Figure 6.7 (bottom) below, so we maintain to use this large time step to reduce the computational cost. For better accuracy, we can locally refine the mesh or trace back more points near the "sharp corner" to avoid the creation of self-intersected trace-back polygons.

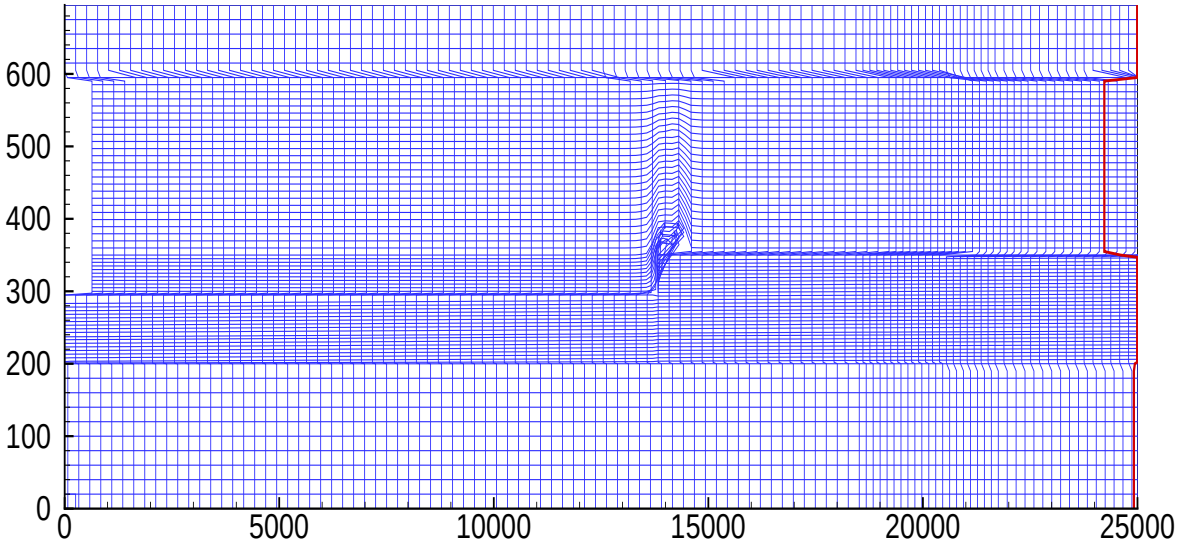


Figure 6.4: Characteristic trace-back mesh. The red polyline near the right edge is the approximation of the trace-forward inflow boundary.

Up to time $3 \times 10^4$ years (Figure 6.5), almost all iodine 129 is still in the clay layer which has a low conductivity and a no-flow boundary condition, so there

is little advection, and the diffusion effect is dominant.

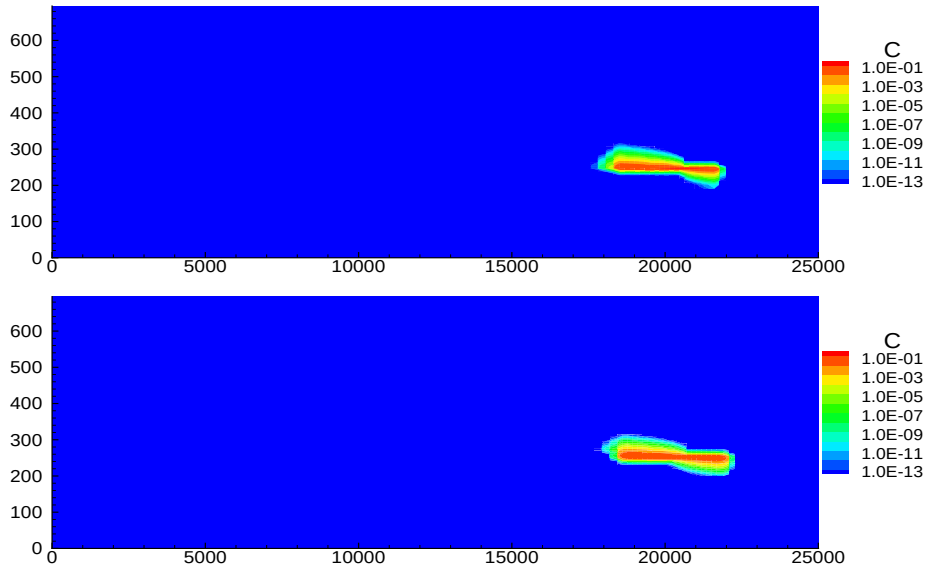


Figure 6.5: Concentrations at $3 \times 10^4$ years approximated by Godunov (top) and VCCMM (bottom).

Figure 6.6 shows the concentration profiles at $2.5 \times 10^5$ years. Due to the relatively high speed in the dogger and limestone layers, the flow front is moving much faster after it escapes from the clay layer. In addition, due to the restricted time step, Godunov's method has more numerical diffusion and shows much wider color bands of concentration contours.

Figure 6.7 shows the concentration profiles at $3 \times 10^5$ years. The flow front is moving much faster in the limestone layer than in the dogger layer due to a much higher conductivity. Each profile shows a sharp concentration jump cross the interface between the limestone and clay layers since the speeds in the two layers have a large difference. There is some inaccuracy of VCCMM approximation near the "sharp corner" of the interface due to the creation of self-intersected trace-back polygons. This inaccuracy should be resolved by a local refinement or tracing back more points.

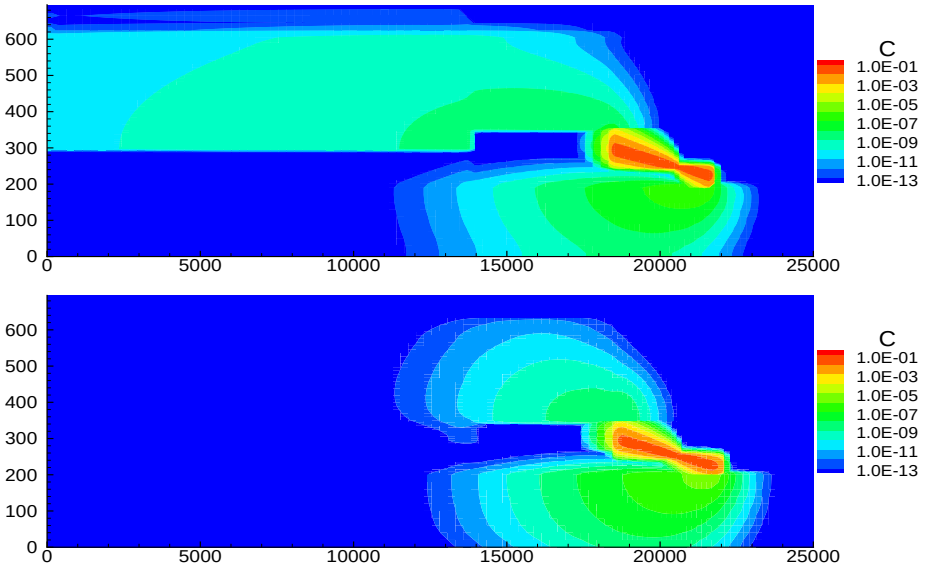


Figure 6.6: Concentrations at $2.5 \times 10^5$ years approximated by Godunov (top) and VCCMM (bottom).

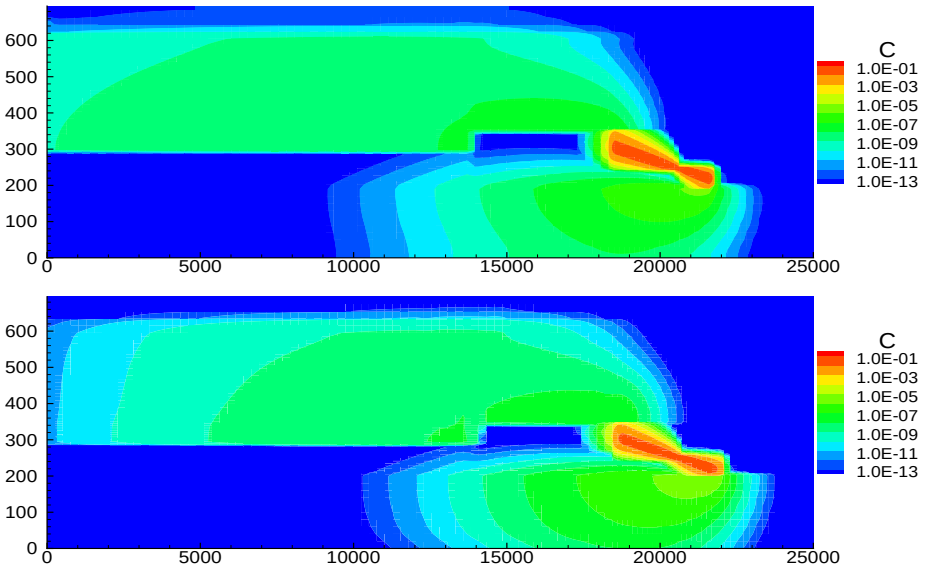


Figure 6.7: Concentrations at $3 \times 10^5$ years approximated by Godunov (top) and VCCMM (bottom).

Figure 6.8 shows the results at $10^6$ years, where Godunov's method and the VCCMM predict a similar shape of the plume.

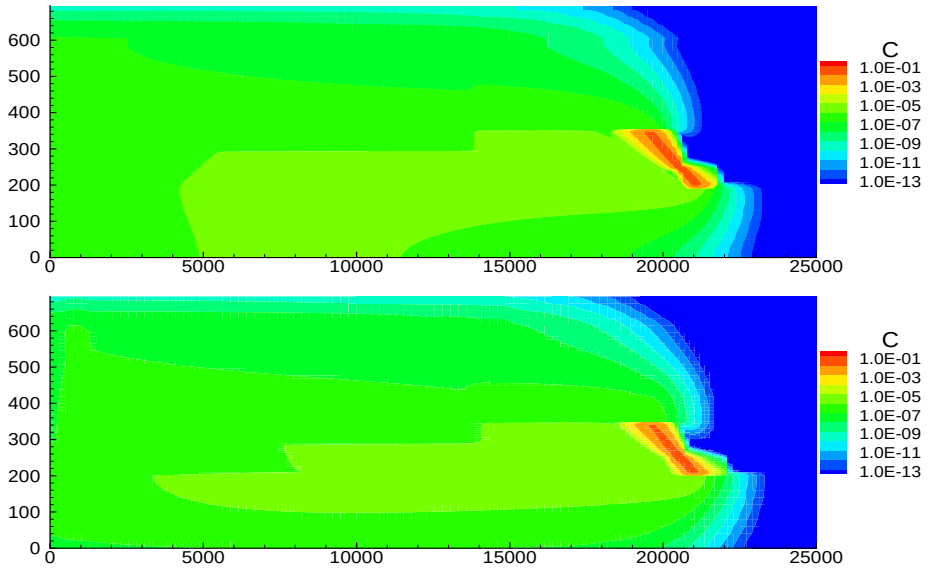


Figure 6.8: Concentrations at $10^6$ years approximated by Godunov (top) and VCCMM (bottom).

A summary of the actual time of computation on a same machine is shown in Table 6.2. Indeed, due to the simple structure of the conductivity distribution, only little time was spent on tracing and volume adjustment, and VCCMM took only about 59.56% of the advection time of Godunov's method[2]. The time differences of diffusion and reaction are due to different solvers.

| | Flow | Trace-back and volume adjustment | Transport | | |
|---|---|---|---|---|---|
| | | | Advection | Diffusion | Reaction |
| Godunov | 0.05 sec | N/A | 1 min 6.77 sec | 20 min 44.9 sec | 0.03 sec |
| VCCMM | 0.05 sec | 2.97 sec | 39.77 sec | 1 min 48.57 sec | 0.15 sec |

Table 6.2: Computational time simulated by VCCMM and Godunov's method.

[2]The approximation with Godunov's method is simulated by Parssim [2].

# Chapter 7

# The Extension to Compressible Flows

The simplest nonlinear variant is to consider compressible problems, in which the fluid density and the medium porosity depend on fluid pressure and may change in time. This model leads to a nonlinear system for the flow. In addition, the local "volume" constraint is then genuinely a *mass* constraint on the bulk fluid.

## 7.1 Flow Approximation

The mass conservation of a compressible bulk fluid gives

$$(\phi\rho)_t + \nabla \cdot (\rho \mathbf{u}) = q \qquad \text{in } \Omega \times J, \tag{7.1}$$

where the porosity $\phi$ and fluid density $\rho$ depend on pressure $p$ which is unknown and may also depend on other physical conditions (e.g., temperature) which are assumed to be given. In general, we denote $\phi = \phi(\mathbf{x}, t, p)$ and $\rho = \rho(\mathbf{x}, t, p)$. The source/sink $q = q(\mathbf{x}, t)$, and the velocity $\mathbf{u}$ is given by the Darcy's law

$$\mathbf{u} = -\mathbf{K}(\nabla p - \rho \mathbf{g}) \qquad \text{in } \Omega \times J, \tag{7.2}$$

where $\mathbf{K} = \mathbf{K}(\mathbf{x})$ is the tensor of medium permeability divided by the fluid viscosity, $p = p(\mathbf{x}, t)$ is the pressure, and $\mathbf{g} \in \mathbb{R}^d$ is the gravitational acceleration. Again, for simplicity, we impose the boundary condition with no flux across $\partial\Omega$, i.e.,

$$\mathbf{u} \cdot \boldsymbol{\nu} = 0 \qquad \text{on } \partial\Omega \times J, \tag{7.3}$$

and an initial condition

$$p(\mathbf{x}, 0) = p^0(\mathbf{x}) \qquad \text{in } \Omega. \tag{7.4}$$

Many techniques have been developed to approximate nonlinear parabolic system (7.1)–(7.2), including discontinuous Galerkin methods [43, 46, 53], linear finite element method [29], and mixed finite element methods [7, 35, 36]. We need to approximate both the density $\rho$ and velocity $\mathbf{u}$, which are also used in the transport approximation. Therefore, a mixed finite element method is employed.

All functions are tacitly assumed to be smooth enough for our purposes. Introduce *effective* fluid density $\tilde{\rho} := \phi\rho$ and *mass* flux rate $\boldsymbol{\psi} := \rho\mathbf{u}$. For notational convenience, we drop the dependence of functions on $\mathbf{x}$ and $t$. Physically, medium porosity $\phi$ and fluid density $\rho$ increase with respect to fluid pressure $p$, so we assume $\tilde{\rho} = \tilde{\rho}(p)$ is strictly increasing in $p$ for any $(\mathbf{x}, t) \in \Omega \times J$, and its inverse function is denoted as $p = p(\tilde{\rho})$. Define

$$f(\tilde{\rho}) := \int_0^{p(\tilde{\rho})} \rho(p)\, dp,$$

$$\boldsymbol{\beta}(f(\tilde{\rho})) := -\mathbf{K} \int_0^{p(\tilde{\rho})} (\nabla_{\mathbf{x}}\rho)(p)\, dp - \rho^2(p(\tilde{\rho}))\mathbf{K}\mathbf{g},$$

where the function $\boldsymbol{\beta}$ is well defined since

$$f_{\tilde{\rho}}(\tilde{\rho}) = \rho(p(\tilde{\rho}))p_{\tilde{\rho}} > 0,$$

which means $f$ is invertible with respect to $\tilde{\rho}$. Then the system (7.1)–(7.4) can be rewritten as

$$\tilde{\rho}_t + \nabla \cdot \boldsymbol{\psi} = q \qquad \text{in } \Omega \times J, \tag{7.5}$$

$$\boldsymbol{\psi} = -\mathbf{K}\nabla f(\tilde{\rho}) - \boldsymbol{\beta}(f(\tilde{\rho})) \qquad \text{in } \Omega \times J, \tag{7.6}$$

$$\boldsymbol{\psi} \cdot \boldsymbol{\nu} = 0 \qquad \text{on } \partial\Omega \times J, \tag{7.7}$$

$$\tilde{\rho}(\mathbf{x}, 0) = \tilde{\rho}^0(\mathbf{x}) \qquad \text{in } \Omega, \tag{7.8}$$

where the initial state is given by

$$\tilde{\rho}^0(\mathbf{x}) := \phi(\mathbf{x}, 0, p^0(\mathbf{x}))\rho(\mathbf{x}, 0, p^0(\mathbf{x})).$$

Following a *nonlinear* mixed finite element method in [7], we impose the following regularity assumptions on functions in system (7.5)–(7.8).

*Assumption* 7.1.1 (Uniformly positive definiteness). Tensor $\mathbf{K} = \mathbf{K}(\mathbf{x})$ is symmetric and uniformly positive definite for $\mathbf{x} \in \Omega$.

*Assumption* 7.1.2 (Non-degeneracy). There exist constants $C_1$ and $C_2$ such that

$$0 < C_1 \leq f_{\tilde{\rho}}(\mathbf{x}, t, \tilde{\rho}) \leq C_2 \qquad \text{for any } (\mathbf{x}, t) \in \Omega \times J \text{ and } \tilde{\rho} \in \mathbb{R}.$$

*Assumption* 7.1.3 (Boundedness of second order derivatives). There exists a constant $C_3$ such that

$$|f_{\tilde{\rho},\tilde{\rho}}(\tilde{\rho})| + |f_{t,\tilde{\rho}}(\tilde{\rho})| \leq C_3 \qquad \text{for any } (\mathbf{x}, t) \in \Omega \times J \text{ and } \tilde{\rho} \in \mathbb{R}.$$

*Assumption* 7.1.4 (Lipschitz continuity). Function $\boldsymbol{\beta}(\varphi) = \boldsymbol{\beta}(\mathbf{x}, t, \varphi)$ is Lipschitz continuous in $L^2$-norm, i.e., there exists some constant $L > 0$, independent of time, such that

$$\|\boldsymbol{\beta}(\varphi_1) - \boldsymbol{\beta}(\varphi_2)\|_2 \leq L\|\varphi_1 - \varphi_2\|_2 \qquad \text{for any } \varphi_1,\ \varphi_2 \in L^2(\Omega).$$

For $(\tilde{\rho}, \boldsymbol{\psi}) = (\tilde{\rho}, \boldsymbol{\psi})(\cdot, t) \in W \times \mathbf{V}$, where $W \subset L^2(\Omega)$ is a scalar function space and $\mathbf{V} \subset \mathbf{H}(\Omega; \mathrm{div})$ is a vector function space with $\boldsymbol{\psi}(\cdot, t) \cdot \boldsymbol{\nu} = 0$ on $\partial\Omega$, the variational form of system (7.5)–(7.8) is

$$(\tilde{\rho}_t, w) + (\nabla \cdot \boldsymbol{\psi}, w) = (q, w), \qquad w \in W, \tag{7.9}$$

$$(\mathbf{K}^{-1}\boldsymbol{\psi}, \mathbf{v}) - (f(\tilde{\rho}), \nabla \cdot \mathbf{v}) + (\mathbf{K}^{-1}\boldsymbol{\beta}(f(\tilde{\rho})), \mathbf{v}) = 0, \quad \mathbf{v} \in \mathbf{V}, \tag{7.10}$$

$$(\tilde{\rho}(\cdot, 0), w) = (\tilde{\rho}^0, w), \qquad w \in W. \tag{7.11}$$

To discretize the time partial derivative and solve this system for $\tilde{\rho}$ and $\boldsymbol{\psi}$, we could use implicit and explicit approximation approaches in time. In terms of

computational effort, the explicit approach is the simplest one at each time step; however, it requires an impractically restricted time step. An efficient and accurate method is a fully implicit approach. The extra cost involved at each time step can be compensated by the fact that larger time steps can be taken.

Let $W_h \subset W$ and $\mathbf{V}_h \subset \mathbf{V}$ be finite element spaces with basis functions $\{w_i\}_{1\le i\le N_W}$ and $\{\mathbf{v}_i\}_{1\le i\le N_V}$, respectively, that satisfy

$$\nabla \cdot \mathbf{V}_h \subset W_h, \tag{7.12}$$

where $\nabla \cdot \mathbf{V}_h := \{\nabla \cdot \mathbf{v}_h : \mathbf{v}_h \in \mathbf{V}_h\}$. For each time step $J^n$, assume the numerical solution $(\tilde{\rho}_h^n, \boldsymbol{\psi}_h^n) \in W_h \times \mathbf{V}_h$. Then the fully discretization of system (7.9)–(7.11) with an implicit marching scheme in time is

$$\left(\frac{\tilde{\rho}_h^{n+1} - \tilde{\rho}_h^n}{\Delta t^n}, w_h\right) - (\nabla \cdot \boldsymbol{\psi}_h^{n+1}, w_h) = (q^{n+1}, w_h), \quad w_h \in W_h, \tag{7.13}$$

$$(\mathbf{K}^{-1}\boldsymbol{\psi}_h^{n+1}, \mathbf{v}_h) - (f(\tilde{\rho}_h^{n+1}), \nabla \cdot \mathbf{v}_h) + (\mathbf{K}^{-1}\boldsymbol{\beta}(f(\tilde{\rho}_h^{n+1})), \mathbf{v}_h) = 0, \quad \mathbf{v}_h \in \mathbf{V}_h, \tag{7.14}$$

$$(\tilde{\rho}_h^0, w_h) = (\tilde{\rho}^0, w_h), \quad w_h \in W_h. \tag{7.15}$$

Let $\tilde{\boldsymbol{\rho}}_h^k \in \mathbb{R}^{N_W}$ be the coefficient vector of $\tilde{\rho}_h^k$ represented as the linear combination of basis $\{w_i\}_{1\le i\le N_W}$, and $\vec{\boldsymbol{\psi}}_h^k \in \mathbb{R}^{N_V}$ be the coefficient vector of $\boldsymbol{\psi}_h^k$ represented as the linear combination of basis $\{\mathbf{v}_i\}_{1\le i\le N_V}$, then the vector form of system (7.13)–(7.15) is

$$\frac{\mathbf{C}(\tilde{\boldsymbol{\rho}}_h^{n+1} - \tilde{\boldsymbol{\rho}}_h^n)}{\Delta t^n} - \mathbf{B}\vec{\boldsymbol{\psi}}_h^{n+1} = \mathbf{q}^{n+1}, \tag{7.16}$$

$$\mathbf{A}\vec{\boldsymbol{\psi}}_h^{n+1} - \mathbf{a}(\tilde{\boldsymbol{\rho}}_h^{n+1}) + \mathbf{b}(\tilde{\boldsymbol{\rho}}_h^{n+1}) = \mathbf{0}, \tag{7.17}$$

$$\mathbf{C}\tilde{\boldsymbol{\rho}}_h^0 = \tilde{\boldsymbol{\rho}}^0, \tag{7.18}$$

where vectors $\mathbf{a} = (a_i)$, $\mathbf{b} = (b_i) \in \mathbb{R}^{N_V}$, $\mathbf{q}^k = (q_i^k)$, $\tilde{\boldsymbol{\rho}}^0 = (\tilde{\rho}_i^0) \in \mathbb{R}^{N_W}$ and matrices $\mathbf{A} = (A_{i,j}) \in \mathbb{R}^{N_V \times N_V}$, $\mathbf{B} = (B_{i,j}) \in \mathbb{R}^{N_W \times N_V}$, $\mathbf{C} = (C_{i,j}) \in \mathbb{R}^{N_W \times N_W}$ are defined

as

$$
\begin{aligned}
&a_i(\tilde{\boldsymbol{\rho}}_h^k) := (f(\tilde{\rho}_h^k), \nabla \cdot \mathbf{v}_i), \quad && b_i(\tilde{\boldsymbol{\rho}}_h^k) := (\mathbf{K}^{-1}\boldsymbol{\beta}(f(\tilde{\rho}_h^k)), \mathbf{v}_i),\\
&q_i^k := (q^k, w_i), && \tilde{\rho}_i^0 := (\tilde{\rho}^0, w_i),\\
&A_{i,j} := (\mathbf{K}^{-1}\mathbf{v}_i, \mathbf{v}_j), && B_{i,j} := (\nabla \cdot \mathbf{v}_j, w_i),\\
&C_{i,j} := (w_i, w_j).
\end{aligned}
$$

We can solve (7.17) for $\vec{\boldsymbol{\psi}}_h^{n+1}$ as

$$\vec{\boldsymbol{\psi}}_h^{n+1} = \mathbf{A}^{-1}(\mathbf{a}(\tilde{\boldsymbol{\rho}}_h^{n+1}) - \mathbf{b}(\tilde{\boldsymbol{\rho}}_h^{n+1})). \tag{7.19}$$

Substituting (7.19) into (7.16) gives

$$\mathbf{f}(\tilde{\boldsymbol{\rho}}_h^{n+1}) = \mathbf{0}, \tag{7.20}$$

where the nonlinear vector-valued function $\mathbf{f}$ is given by

$$\mathbf{f}(\mathbf{p}_h^{n+1}) := \frac{\mathbf{C}(\tilde{\boldsymbol{\rho}}_h^{n+1} - \tilde{\boldsymbol{\rho}}_h^n)}{\Delta t^n} - \mathbf{B}\mathbf{A}^{-1}(\mathbf{a}(\tilde{\boldsymbol{\rho}}_h^{n+1}) - \mathbf{b}(\tilde{\boldsymbol{\rho}}_h^{n+1})) - \mathbf{q}^{n+1}. \tag{7.21}$$

Nonlinear equation (7.20) is expected to be solved for $\tilde{\boldsymbol{\rho}}_h^{n+1}$ by iterative methods such as Newton's method, then $\vec{\boldsymbol{\psi}}_h^{n+1}$ is computed by (7.19).

If the fluid and porous medium are *slightly* compressible, in practice, we assume that the density $\rho$ and the porosity $\phi$ are linearly dependent on the pressure [19, pp. 15]. That is

$$\rho(p) = \rho_{\text{ref}}(1 + c_f(p - p_{\text{ref}})) \quad \text{and} \quad \phi(p) = \phi_{\text{ref}}(1 + c_r(p - p_{\text{ref}})), \tag{7.22}$$

where $\rho_{\text{ref}}$ and $\phi_{\text{ref}}$ are the reference values at the reference pressure $p_{\text{ref}}$, and $c_f$ and $c_r$ are the fluid compressibility and the rock compressibility, respectively, assumed constant. In this case, equation (7.21) could be simplified, leading to a simplified nonlinear equation (7.20).

An optimal $L^2$-error estimate of the fully discrete mixed finite element approximation (7.13)–(7.15) is stated in the following theorem [7].

**Theorem 7.1.1** (Optimal error estimate)**.** *Let Assumptions* 7.1.1–7.1.4 *hold and time grid* $\{t^n\}_{0\le n\le N}$ *is regular as in Assumption* 3.2.3*. Assume* $(\tilde{\rho}, \boldsymbol{\psi}) \in W \times \mathbf{V}$ *is smooth enough which solves system* (7.5)–(7.8)*, and* $(\tilde{\rho}_h^k, \boldsymbol{\psi}_h^k) \in W_h \times \mathbf{V}_h$*, where* $W_h$ *and* $\mathbf{V}_h$ *satisfy* (7.12)*, solves its fully discrete mixed finite element approximation* (7.13)–(7.15)*. Then there is some constant* $C > 0$ *such that if* $\Delta t$ *is sufficiently small, then for* $n$ *between* 1 *and* $N$*,*

$$\|\tilde{\rho}_h^n - \tilde{\rho}^n\|_2^2 + \sum_{k=1}^{n} \|\boldsymbol{\psi}_h^k - \boldsymbol{\psi}^k\|_2^2 \Delta t^{k-1}$$
$$\le C \left\{ \|P_h\tilde{\rho}^0 - \tilde{\rho}^0\|_2^2 + (\Delta t)^2 + \int_0^{t^n} \|P_h\tilde{\rho}_t - \tilde{\rho}_t\|_2^2 \, dt + \sum_{k=1}^{n} \|\Pi_h\boldsymbol{\psi}^k - \boldsymbol{\psi}^k\|_2^2 \Delta t^{k-1} \right\},$$

*where* $P_h$ *is the* $L^2$*-projection operator associated with space* $W_h$*, and for any* $\mathbf{v} \in \mathbf{H}(\Omega; \mathrm{div})$*, define* $\Pi_h\mathbf{v} \in \mathbf{V}_h$*, such that* $\|\Pi_h\mathbf{v} - \mathbf{v}\|_2$ *is minimal subject to the constraint* $\nabla \cdot \Pi_h\mathbf{v} = P_h\nabla \cdot \mathbf{v}$*.*

Let $\tilde{\rho}_h$ and $\boldsymbol{\psi}_h$ be some interpolations of $\tilde{\rho}_h^n$ and $\boldsymbol{\psi}_h^n$ in time $J_T$, respectively, such that new local extremas are not introduced (e.g., linear interpolation). Let $\mathbf{v}_h := \boldsymbol{\psi}_h/\tilde{\rho}_h$ be the approximation of *interstitial* velocity $\mathbf{v} := \boldsymbol{\psi}/\tilde{\rho}$. In general, for our purpose, denote a total $L^\infty$-error $\varepsilon_{\text{flow}}$ due to the flow approximation

$$\varepsilon_{\text{flow}} := \|\tilde{\rho}_h - \tilde{\rho}\|_\infty + \|\mathbf{v}_h - \mathbf{v}\|_\infty + \|\nabla \cdot (\mathbf{v}_h - \mathbf{v})\|_\infty, \tag{7.23}$$

which is used in the convergence analysis in Section 7.4.

## 7.2 Transport Approximation

The mass conservation of tracer gives

$$(\phi\rho c)_t + \nabla \cdot (\rho c\mathbf{u}) = q_c := c_I q^+ + cq^- \qquad \text{in } \Omega \times J, \tag{7.24}$$

In terms of *effective* fluid density $\tilde{\rho} = \phi\rho$ and *interstitial* velocity $\mathbf{v} = \boldsymbol{\psi}/\tilde{\rho}$, it can be rewritten as

$$(\tilde{\rho}c)_t + \nabla \cdot (\tilde{\rho}c\mathbf{v}) = q_c, \tag{7.25}$$

which has a similar form to the transport equation (2.2) of incompressible fluid. So we can treat (7.25) in the same way except that the effective fluid density $\tilde{\rho}$ is now dependent on time. Then the local volume constraint (2.15) is now genuinely a *mass* constraint on the bulk fluid as shown in (7.26) below. A description of the algorithm of transport approximation follows.

**Algorithm for Transport Approximation**

(VCCMM for Compressible Flows)

**Step 1: Form velocity field.** For each time step $J^n$, form a velocity field $\mathbf{v}_h$ by interpolating *interstitial* velocity $\mathbf{v}_h^n := \boldsymbol{\psi}_h^n/\tilde{\rho}_h^n$ and $\mathbf{v}_h^{n+1} := \boldsymbol{\psi}_h^{n+1}/\tilde{\rho}_h^{n+1}$ in time.

**Step 2: Compute trace-backs/trace-forwards.** Compute trace-back or trace-forward regions by tracing each grid element $E$ with velocity field $\mathbf{v}_h$ to a region $\tilde{E}_h^n$ approximated by a polygon. (The subscript $h$ of trace-back regions means tracing back with velocity $\mathbf{v}_h$.)

**Step 3: Trace-back adjustment.** Adjust regions $\tilde{E}_h^n$ as described in the Volume Correction Algorithm to obtain the *local mass conservation of bulk fluid*

$$M_h^{n+1}(E) = M_h^n(\tilde{E}_h^n) + \iint_{\tilde{\mathcal{E}}_h^n} q \, d\mathbf{x} \, dt, \tag{7.26}$$

where $M_h^k(S)$ is the numerical mass of the bulk fluid in a region $S \subset \Omega$ at time $t^k$ defined as

$$M_h^k(S) := \int_S \tilde{\rho}_h^k \, d\mathbf{x}.$$

**Step 4: Update numerical solution.** By the *local mass conservation of the tracer*, $c_h^{n+1} \in W_h(\Omega)$ is defined on $E$ to be

$$c_{h,E}^{n+1} M_h^{n+1}(E) = \int_{\tilde{E}_h^n} \tilde{\rho}_h^n c_h^n \, d\mathbf{x} + \iint_{\tilde{\mathcal{E}}_h^n} q_{c_h^n} \, d\mathbf{x} \, dt. \tag{7.27}$$

By the design in (7.26) and (7.27), this method is locally conservative for mass of *both* bulk fluid and tracer (i.e., a *fully* conservative method).

## 7.3 Stability Analysis

We continue to use notations $\mathbf{c}_h^n := (c_{h,E}^n)_{E\in\mathcal{T}_h} \in \mathbb{R}^{N_h}$ for $0 \le n \le N$ and $\mathcal{T}_{h,P}$ (i.e., the production wells) as in Section 2.3. Similar to the incompressible case, we also have the scheme (7.27) of VCCMM for compressible flows in a vector form. Since the density of compressible flows may change in time, we should measure the amount of fluid in a region by *mass* instead of volumes. Replacing volumes by numerical mass in (2.23) and (2.24), we have

$$A_{h,E,F}^n := \begin{cases} \dfrac{M_h^n(\tilde{E}_h^n \cap F)}{M_h^{n+1}(E)}, & E \notin \mathcal{T}_{h,P}, \\ \dfrac{M_h^n(\tilde{E}_h^n \cap F)}{M_h^{n+1}(E)} \left(1 + \dfrac{(M_E^n)^-}{M_h^n(\tilde{E}_h^n \setminus E)}\right), & E \in \mathcal{T}_{h,P},\ F \ne E, \\ \dfrac{(M_E^n)^+}{M_h^{n+1}(E)}, & E = F \in \mathcal{T}_{h,P}, \end{cases} \tag{7.28}$$

and

$$b_{h,E}^n := \frac{1}{M_h^{n+1}(E)} \iint_{\tilde{\mathcal{E}}_{E,h}^n} c_I q^+ \, d\mathbf{x}\, dt, \tag{7.29}$$

where $M_E^n$ is the remaining mass of the bulk fluid in the production well in $E$ at time $t^n$ defined as

$$M_E^n := M_h^n(E) + \iint_{I_E^n} q^- \, d\mathbf{x}\, dt. \tag{7.30}$$

Then the the scheme of VCCMM for compressible flows in a vector form is given by

$$\mathbf{c}_h^{n+1} = \mathbf{A}_h^n \mathbf{c}_h^n + \mathbf{b}_h^n, \tag{7.31}$$

where $\mathbf{A}_h^n = (A_{h,E,F}^n) \in \mathbb{R}^{N_h \times N_h}$ and $\mathbf{b}_h^n = (b_{h,E}^n) \in \mathbb{R}^{N_h}$.

The following lemma shows that matrix $\mathbf{A}_h^n$ has the same property as in Lemma 2.3.1.

**Lemma 7.3.1.** *The matrix $\mathbf{A}_h^n$ defined in* (7.28) *as a vector operator does not increase the $l^\infty$-norm of a vector, i.e., for any $\mathbf{c} \in \mathbb{R}^{N_h}$,*

$$|\mathbf{A}_h^n \mathbf{c}|_\infty \le |\mathbf{c}|_\infty.$$

*Proof.* First we will show each entry $A^n_{h,E,F} \geq 0$. By (7.28), we only need to show

$$M^n_h(\tilde{E}^n_h \setminus E) + (M^n_E)^- \geq 0 \tag{7.32}$$

for $E \in \mathcal{T}_{h,P}$ and $F \neq E$. Actually, if $M^n_E \geq 0$, it is trivial that inequality (7.32) holds. Otherwise,

$$\begin{aligned} M^n_h(\tilde{E}^n_h \setminus E) + (M^n_E)^- &= M^n_h(\tilde{E}^n_h \setminus E) + M^n_E \\ &= M^n_h(\tilde{E}^n_h) - M^n_h(E) + \left( M^n_h(E) + \iint_{I^n_E} q^- \, d\mathbf{x}\, dt \right) \\ &= M^n_h(\tilde{E}^n_h) + \iint_{\tilde{\mathcal{E}}^n_{E,h}} q \, d\mathbf{x}\, dt = M^{n+1}_h(E) \geq 0, \end{aligned}$$

where the last equality is obtained by (7.26).

Now we show each row sum of $\mathbf{A}^n_h$

$$\sum_{F \in \mathcal{T}_h} A^n_{h,E,F} \leq 1. \tag{7.33}$$

By (7.28) and (7.26), we compute when $E \notin \mathcal{T}_{h,P}$,

$$\sum_{F \in \mathcal{T}_h} A^n_{h,E,F} = \frac{M^n_h(\tilde{E}^n_h)}{M^{n+1}_h(E)} = \frac{1}{M^{n+1}_h(E)} \left( M^{n+1}_h(E) - \iint_{\tilde{\mathcal{E}}^n_{E,h}} q \, d\mathbf{x}\, dt \right) \leq 1.$$

and when $E \in \mathcal{T}_{h,P}$,

$$\begin{aligned} \sum_{F \in \mathcal{T}_h} A^n_{h,E,F} &= \frac{M^n_h(\tilde{E}^n_h \setminus E)}{M^{n+1}_h(E)} \left( 1 + \frac{(M^n_E)^-}{M^n_h(\tilde{E}^n_h \setminus E)} \right) + \frac{(M^n_E)^+}{M^{n+1}_h(E)} \\ &= \frac{M^n_h(\tilde{E}^n_h \setminus E) + M^n_E}{M^{n+1}_h(E)} = \frac{1}{M^{n+1}_h(E)} \left( M^n_h(\tilde{E}^n_h) + \iint_{I^n_E} q^- \, d\mathbf{x}\, dt \right) \\ &= \frac{1}{M^{n+1}_h(E)} \left( M^n_h(\tilde{E}^n_h) + \iint_{\tilde{\mathcal{E}}^n_{E,h}} q \, d\mathbf{x}\, dt \right) = 1. \end{aligned}$$

So we obtain (7.33). Then the same argument as in Lemma 2.3.1 is performed to complete the proof. □

By Lemma 7.3.1 and using the same argument in Theorem 2.3.2, we obtain the stability of VCCMM for compressible flows stated as following.

**Theorem 7.3.2** (Stability of VCCMM for compressible flows)**.** *The scheme of VCCMM for compressible flows given by* (7.31) *is stable. That is, if* $\mathbf{c}_h^n$ $(0 \leq n \leq N)$ *satisfies scheme* (7.31) *in time* $J_T$ *with an initial approximation* $\mathbf{c}_h^0$*, and* $\tilde{\mathbf{c}}_h^n$ $(0 \leq n \leq N)$ *satisfies the perturbed scheme*

$$\tilde{\mathbf{c}}_h^{n+1} = \mathbf{A}_h^n \tilde{\mathbf{c}}_h^n + \mathbf{b}_h^n + \Delta t^n \boldsymbol{\delta}_h^n$$

*in time* $J_T$ *with an initial approximation* $\tilde{\mathbf{c}}_h^0$*, where* $\boldsymbol{\delta}_h^n \in \mathbb{R}^{N_h}$ *is a perturbation at time step* $J^n$*, then the following error estimate holds:*

$$\max_{0 \leq n \leq N} |\tilde{\mathbf{c}}_h^n - \mathbf{c}_h^n|_\infty \leq |\tilde{\mathbf{c}}_h^0 - \mathbf{c}_h^0|_\infty + T \max_{0 \leq n \leq N} |\boldsymbol{\delta}_h^n|_\infty.$$

## 7.4 Convergence Analysis

Again, we use the key idea introduced by Arbogast and Wheeler [5], and construct a perturbed velocity $\tilde{\mathbf{v}}_h$ such that, for each time step $J^n$, each trace-back element $\tilde{E}_h^n$ satisfies the local mass conservation of bulk fluid (7.26), and the numerical solution $c_h$ weakly satisfies the perturbed system

$$(\tilde{\rho}_h c_h)_t + \nabla \cdot (\tilde{\rho}_h c_h \tilde{\mathbf{v}}_h) = q_{c_h} \quad \text{in } \Omega \times J^n, \tag{7.34}$$

$$c_h(\mathbf{x}, t^n) = c_h^n(\mathbf{x}) \quad \text{in } \Omega. \tag{7.35}$$

Then the update $c_h^{n+1}$ is defined as

$$c_h^{n+1}(\mathbf{x}) := \tilde{P}_h^{n+1} c_h(\mathbf{x}, t^{n+1}-) = \tilde{P}_h^{n+1} c_h^{n+1-}(\mathbf{x}), \tag{7.36}$$

where the weighted $L^2$-projection operator $\tilde{P}_h^k$ is defined as

$$(\tilde{P}_h^k f, w)_{k,h} = (f, w)_{k,h} \text{ for all } w \in W_h(\Omega) \tag{7.37}$$

with the weighted $L^2$-inner product given by

$$(\varphi_1, \varphi_2)_{k,h} := \int_\Omega \varphi_1 \varphi_2 \tilde{\rho}_h^k \, d\mathbf{x} \text{ for all } \varphi_1,\ \varphi_2 \in L^2(\Omega).$$

Similar to Assumption 3.0.1, we first make the following assumption of the existence and an error estimate of the perturbed velocity field $\tilde{\mathbf{v}}_h$ required by (7.26) and system (7.34)–(7.35). The proof of this assumption is given in Section 7.4.2.

*Assumption* 7.4.1 (Perturbed velocity field). The velocity field $\mathbf{v} \in C^1(\Omega \times J_T)$ has divergence $\nabla \cdot \mathbf{v}(\cdot, t)$ uniformly Lipschitz continuous in time $J_T$, i.e.,

$$|\nabla \cdot \mathbf{v}(\mathbf{x}, t) - \nabla \cdot \mathbf{v}(\mathbf{y}, t)| \leq L|\mathbf{x} - \mathbf{y}| \quad \text{for all } \mathbf{x},\ \mathbf{y} \in \Omega,\ t \in J_T,$$

where $L > 0$ is a constant independent of $\mathbf{x}$, $\mathbf{y}$, and $t$. There exists a locally conservative velocity field $\tilde{\mathbf{v}}_h = \tilde{\mathbf{v}}_h(\mathbf{x}, t)$ on $\Omega \times J_T$ such that

$$\tilde{\mathbf{v}}_h \cdot \boldsymbol{\nu} = 0 \text{ on } \partial\Omega \times J_T,$$

each trace-back polygon $\tilde{E}_h^n$ satisfies the local mass constraint (7.26), and

$$\|\mathbf{v} - \tilde{\mathbf{v}}_h\|_\infty + \|\nabla \cdot (\mathbf{v} - \tilde{\mathbf{v}}_h)\|_\infty \leq C \left( h + (\Delta t)^r + \big(\frac{h}{\Delta t} + 1\big)\varepsilon_{\text{flow}} \right),$$

where $C$ and $r > 0$ are constants independent of $h$ and $\Delta t$, and $\varepsilon_{\text{flow}}$ is given by (7.23).

### 7.4.1 Convergence results

Introduce the *effective* concentration of tracer $\tilde{c} := \tilde{\rho} c$ and the numerical *effective* concentration of tracer $\tilde{c}_h := \tilde{\rho}_h c_h$, then $\tilde{c}$ solves

$$\tilde{c}_t + \nabla \cdot (\tilde{c}\mathbf{v}) = c_I q^+ + \tilde{c}\tilde{q}^- \quad \text{in } \Omega \times J_T, \tag{7.38}$$

and in each $J^n$, $\tilde{c}_h$ solves

$$(\tilde{c}_h)_t + \nabla \cdot (\tilde{c}_h \tilde{\mathbf{v}}_h) = c_I q^+ + \tilde{c}_h \tilde{q}_h^- \quad \text{in } \Omega \times J^n, \tag{7.39}$$

where $\tilde{\mathbf{v}}_h$ is the perturbed velocity field in Assumption 7.4.1, $\tilde{q} := q/\tilde{\rho}$ and $\tilde{q}_h := q/\tilde{\rho}_h$. Then (7.38) and (7.39) have the same form as the incompressible transport equation (2.8). The convergence result for incompressible flows in Theorem 3.4.1 is based on Lemmas 3.3.1–3.3.3, and we extend these lemmas to compressible flows.

Let all assumptions in Theorem 3.4.1 hold for $\tilde{c}$, $\tilde{c}_h$, and $\mathbf{v}$. Similar to (3.32), for $\tilde{c}$ and $\tilde{c}_h$, define

$$\tilde{\rho}^n_{\varepsilon,h} := \iint_{\Omega\times\Omega} K_\varepsilon(\mathbf{x}-\mathbf{y})|\tilde{c}^n(\mathbf{x}) - \tilde{c}^n_h(\mathbf{y})|\, d\mathbf{x}d\mathbf{y}, \tag{7.40}$$

where $K_\varepsilon$ is an approximation of the identity in $\Omega$, i.e.,

$$K_\varepsilon(\mathbf{x}) := \frac{1}{\varepsilon^d} K_0\left(\frac{\mathbf{x}}{\varepsilon}\right), \qquad \varepsilon > 0,$$

where function $K_0$ is non-negative, smooth and compactly supported in $\Omega$ with an integral of one.

**Extension of Lemma 3.3.1.** By definition of $\tilde{\rho}^n_{\varepsilon,h}$ in (7.40), it is trivial to prove that Lemma 3.3.1 still holds for $\tilde{\rho}^n_{\varepsilon,h}$, i.e.,

$$|\, \tilde{\rho}^n_{\varepsilon,h} - \|\tilde{c}^n - \tilde{c}^n_h\|_1 \,| \le C\varepsilon. \tag{7.41}$$

**Extension of Lemma 3.3.2.** By (7.38) and (7.39), $\tilde{c}$ and $\tilde{c}_h$ can be treated as $c$ and $c_h$, respectively, in the proof of Lemma 3.3.2 except that (7.38) and (7.39) have different coefficient functions, $\tilde{q}$ and $\tilde{q}_h$, respectively. By (7.23), we have

$$\|\tilde{q}_h - \tilde{q}\|_\infty = \|q/\tilde{\rho}_h - q/\tilde{\rho}\|_\infty \le \frac{\|q\|_\infty}{\tilde{\rho}^2_*}\|\tilde{\rho}_h - \tilde{\rho}\|_\infty \le C\varepsilon_{\text{flow}}. \tag{7.42}$$

By (7.42), Assumption 7.4.1 and applying the argument in Lemma 3.3.2 to $\tilde{\rho}^n_{\varepsilon,h}$,

$$\tilde{\rho}^{n+1-}_{\varepsilon,h} - \tilde{\rho}^n_{\varepsilon,h} \le C\Delta t\big(\varepsilon + h + (\Delta t)^r + (h/\Delta t + 1)\varepsilon_{\text{flow}}\big). \tag{7.43}$$

**Extension of Lemma 3.3.3.** In this section, we only focus on the case $W_h = W_h(\Omega)$ in the flow approximation of mixed finite element method (7.13)–(7.15),

which includes $\mathrm{RT}_0$ and $\mathrm{AW}_0$ spaces. Then $\tilde{\rho}_h^k \in W_h(\Omega)$, and the weighted $L^2$-projection operator $\tilde{P}_h^k$ defined in (7.37) coincides with $P_h$. Multiplying $\tilde{\rho}_h^{n+1}$ on both sides of (7.36), we have

$$\tilde{c}_h^{n+1} = P_h \tilde{c}_h^{n+1-},$$

which has the same form as (3.9). So applying the argument in Lemma 3.3.3 to $\tilde{\rho}_{\varepsilon,h}^n$ gives

$$\tilde{\rho}_{\varepsilon,h}^n - \tilde{\rho}_{\varepsilon,h}^{n-} \le C\frac{h^2}{\varepsilon}|\tilde{c}_h^{n-}|_{BV}. \tag{7.44}$$

**Theorem 7.4.1** (Convergence of VCCMM for compressible flows)**.** *Let all assumptions in Theorem* 3.4.1 *hold for* $\tilde{c}$, $\tilde{c}_h$, *and* $\mathbf{v}$, *and let Assumption* 7.4.1 *hold. Then the following* $L^1$*-error estimate holds:*

$$\max_{0\le n\le N}\|c_h^n - c^n\|_1 \le C\left(\|c_h^0 - c^0\|_1 + \frac{h}{\sqrt{\Delta t}} + h + (\Delta t)^r + \big(\frac{h}{\Delta t}+1\big)\varepsilon_{\text{flow}}\right), \tag{7.45}$$

*where* $C$ *and* $r > 0$ *are constants independent of* $h$ *and* $\Delta t$.

*Proof.* By (7.41), (7.43) and (7.44), applying the argument in Theorem 3.4.1 to $\tilde{c}$ and $\tilde{c}_h$, we have

$$\max_{0\le n\le N}\|\tilde{c}_h^n - \tilde{c}^n\|_1 \le \|\tilde{c}_h^0 - \tilde{c}^0\|_1 + C\left(\frac{h}{\sqrt{\Delta t}} + h + (\Delta t)^r + \big(\frac{h}{\Delta t}+1\big)\varepsilon_{\text{flow}}\right). \tag{7.46}$$

Since $\tilde{c} = \tilde{\rho}c$ and $\tilde{c}_h^n = \tilde{\rho}_h^n c_h^n$, we have

$$\begin{aligned}\|c_h^n - c^n\|_1 &= \|\tilde{c}_h^n/\tilde{\rho}_h^n - \tilde{c}^n/\tilde{\rho}^n\|_1 \le \tilde{\rho}_*^{-2}\|\tilde{\rho}^n\tilde{c}_h^n - \tilde{\rho}_h^n\tilde{c}^n\|_1\\ &\le \tilde{\rho}_*^{-2}(\|\tilde{\rho}^n\|_\infty\|\tilde{c}_h^n - \tilde{c}^n\|_1 + \|\tilde{c}^n\|_\infty\|\tilde{\rho}_h^n - \tilde{\rho}^n\|_1)\\ &\le C(\|\tilde{c}_h^n - \tilde{c}^n\|_1 + \varepsilon_{\text{flow}}),\end{aligned} \tag{7.47}$$

and

$$\begin{aligned}\|\tilde{c}_h^0 - \tilde{c}^0\|_1 &= \|\tilde{\rho}_h^0 c_h^0 - \tilde{\rho}^0 c^0\|_1 \le \|c_h^0\|_\infty\|\tilde{\rho}_h^0 - \tilde{\rho}^0\|_1 + \|\tilde{\rho}^0\|_\infty\|c_h^0 - c^0\|_1\\ &\le C(\varepsilon_{\text{flow}} + \|c_h^0 - c^0\|_1).\end{aligned} \tag{7.48}$$

Combining (7.46), (7.47) and (7.48), we obtain (7.45) and complete the proof. □

*Remark* 7.4.1. The error in (7.45) is consistent with the result of incompressible flows in (3.40). The only extra term $\mathcal{O}((h/\Delta t + 1)\varepsilon_{\text{flow}})$ in the error estimate of compressible flows is contributed from the perturbation of velocity field in Assumption 7.4.1, and is purely due to the flow approximation.

### 7.4.2 Perturbed velocity field

In this section, we construct the perturbed velocity $\tilde{\mathbf{v}}_h$ and prove Assumption 7.4.1. Note that the error $\mathbf{v} - \tilde{\mathbf{v}}_h$ can be written as

$$\mathbf{v} - \tilde{\mathbf{v}}_h = (\mathbf{v} - \mathbf{v}_h) + (\mathbf{v}_h - \tilde{\mathbf{v}}_h),$$

where the first term is the error due to flow approximation included in $\varepsilon_{\text{flow}}$ given by (7.23), and the estimate of the second term is almost identical to the incompressible case in Assumption 3.0.1 except that now we measure the bulk fluid in mass rather than volumes. The proof of Assumption 3.0.1 consists Lemmas 3.5.1, 3.5.2, 3.5.4, and 3.5.5. Lemmas 3.5.1 and 3.5.4 give the construction of the perturbed velocity, and we still use this construction for compressible flows by replacing $\mathbf{u}$ with $\mathbf{v}_h$. Since $\mathbf{v}_h$ is the numerical solution from the flow approximation, it is very likely that we lose some regularity of $\mathbf{v}_h$. However, in practice, we can always adjust points by this construction without verifying regularity of $\mathbf{v}_h$ as long as the perturbation is sufficiently small. So we tacitly assume that the conclusions of Lemmas 3.5.1 and 3.5.4 still hold for $\mathbf{v}_h$. Lemmas 3.5.2 and 3.5.5 deal with trace-back adjustment which involves volumes of trace-back regions, so we need to replace these volumes with mass and give similar proofs.

In each time step $J^n$, let $\check{R}^n$ and $\check{R}_h^n$ be the exact trace-back regions of ring $R$ with velocities $\mathbf{v}$ and $\mathbf{v}_h$, respectively. Let $\tilde{R}_h^n(\alpha)$ be the trace-back polygon of ring $R$ shown in Figure 3.1. We assume characteristic tracing is exact for the moment. In addition to Assumptions 3.5.1–3.5.4, we make another assumption on $\tilde{\rho}_h^k$.

*Assumption* 7.4.2 (Uniform boundedness). The effective density $\tilde{\rho}$ and its mixed finite element approximation $\tilde{\rho}_h$ are uniformly bounded away from zero, i.e., there exist constants $\tilde{\rho}_*$ and $\tilde{\rho}^*$ such that

$$0 < \tilde{\rho}_* \le \tilde{\rho}(\mathbf{x}, t) \le \tilde{\rho}^* \quad \text{and} \quad 0 < \tilde{\rho}_* \le \tilde{\rho}_h(\mathbf{x}, t) \le \tilde{\rho}^*$$

for any $(\mathbf{x}, t) \in \Omega \times J_T$ and $h > 0$.

The following lemma gives the estimate of trace-back errors due to flow approximation.

**Lemma 7.4.2** (Trace-back error)**.** *Let $\check{\mathbf{x}}^n$ and $\check{\mathbf{x}}_h^n$ are the trace-back points of $\mathbf{x}$ from time $t^{n+1}$ to $t^n$ with velocities $\mathbf{v}$ and $\mathbf{v}_h$, respectively. Then the error*

$$|\check{\mathbf{x}}^n - \check{\mathbf{x}}_h^n| \le C\varepsilon_{\text{flow}}\Delta t^n, \tag{7.49}$$

*where $C > 0$ is a constant independent of $n$, $h$, and $\Delta t^n$.*

*Proof.* Let $C_* := (1 - t_*\|\nabla \mathbf{v}\|_\infty)^{-1} > 1$, where we fix some $t_* > 0$ such that $t_*\|\nabla \mathbf{v}\|_\infty < 1$. Let $\mathbf{e}_h(t) := \check{\mathbf{x}}(t) - \check{\mathbf{x}}_h(t)$ be the trace-back error. By induction, we show that for any integer $k \ge 1$ and $0 < t \le kt_*$,

$$\|\mathbf{e}_h\|_{\infty,[t^{n+1}-t,t^{n+1}]} \le C_*^k \varepsilon_{\text{flow}} t. \tag{7.50}$$

Let $k = 1$ and $0 < t \le t_*$. For any $s \in [t^{n+1} - t, t^{n+1}]$, note that $\mathbf{e}_h(t^{n+1}) = \mathbf{0}$, and we have

$$\begin{aligned}
|\mathbf{e}_h(s)| &\le |e_h'(\xi)|(t^{n+1} - s) \\
&= |\mathbf{v}(\check{\mathbf{x}}(\xi), \xi) - \mathbf{v}_h(\check{\mathbf{x}}_h(\xi), \xi)|(t^{n+1} - s) \\
&\le |\mathbf{v}(\check{\mathbf{x}}(\xi), \xi) - \mathbf{v}(\check{\mathbf{x}}_h(\xi), \xi)|\, t + |\mathbf{v}(\check{\mathbf{x}}_h(\xi), \xi) - \mathbf{v}_h(\check{\mathbf{x}}_h(\xi), \xi)|\, t \\
&\le \|\nabla \mathbf{v}\|_\infty \|\mathbf{e}_h\|_{\infty,[t^{n+1}-t,t^{n+1}]} t_* + \varepsilon_{\text{flow}} t,
\end{aligned}$$

where $\xi \in (s, t^{n+1})$ comes from the mean value theorem. Since $s \in [t^{n+1} - t, t^{n+1}]$ is arbitrary,

$$\|\mathbf{e}_h\|_{\infty,[t^{n+1}-t,t^{n+1}]} \le \|\nabla \mathbf{v}\|_\infty \|\mathbf{e}_h\|_{\infty,[t^{n+1}-t,t^{n+1}]} t_* + \varepsilon_{\text{flow}} t,$$

which gives (7.50) with $k = 1$.

Assume (7.50) holds for some integer $k$. For $kt_* < t \le (k+1)t_*$ and any $s \in [t^{n+1} - t, t^{n+1} - kt_*]$,

$$\begin{aligned}
|\mathbf{e}_h(s)| &\le |e_h(t^{n+1} - kt_*)| + |e_h(s) - e_h(t^{n+1} - kt_*)| \\
&\le C_*^k \varepsilon_{\text{flow}} kt_* + |e_h'(\xi)|(t^{n+1} - kt_* - s) \\
&= C_*^k \varepsilon_{\text{flow}} kt_* + |\mathbf{v}(\check{\mathbf{x}}(\xi), \xi) - \mathbf{v}_h(\check{\mathbf{x}}_h(\xi), \xi)|(t^{n+1} - kt_* - s) \\
&\le C_*^k \varepsilon_{\text{flow}} kt_* + |\mathbf{v}(\check{\mathbf{x}}(\xi), \xi) - \mathbf{v}(\check{\mathbf{x}}_h(\xi), \xi)|t_* + |\mathbf{v}(\check{\mathbf{x}}_h(\xi), \xi) - \mathbf{v}_h(\check{\mathbf{x}}_h(\xi), \xi)|(t - kt_*) \\
&\le C_*^k \varepsilon_{\text{flow}} t + \|\nabla \mathbf{v}\|_\infty \|\mathbf{e}_h\|_{\infty, [t^{n+1}-t, t^{n+1}]} t_*,
\end{aligned}$$

where $\xi \in (s, t^{n+1} - kt_*)$ comes from the mean value theorem. Since $s \in [t^{n+1} - t, t^{n+1} - kt_*]$ is arbitrary,

$$\begin{aligned}
\|\mathbf{e}_h\|_{\infty, [t^{n+1}-t, t^{n+1}]} &= \max\{\|\mathbf{e}_h\|_{\infty, [t^{n+1}-kt_*, t^{n+1}]}, \|\mathbf{e}_h\|_{\infty, [t^{n+1}-t, t^{n+1}-kt_*]}\} \\
&\le C_*^k \varepsilon_{\text{flow}} t + \|\nabla \mathbf{v}\|_\infty \|\mathbf{e}_h\|_{\infty, [t^{n+1}-t, t^{n+1}]} t_*,
\end{aligned}$$

which gives (7.50) with $k$ replaced by $k+1$. So (7.50) holds for any integer $k \ge 1$. Taking $t = \Delta t^n$, we obtain (7.49) with $C = C_*^{k_*}$, where $k_* := \min\{k \text{ integer} : kt_* \ge T\}$. □

**Lemma 7.4.3** (Change rate of mass for compressible flows)**.** *Let $m_h^n(\alpha) := M_h^n(\tilde{R}_h^n(\alpha))$ be the numerical mass of the adjusted trace-back ring, and assume $m_h^n(\alpha)$ is differentiable with respect to $\alpha$. Let Assumptions* 3.5.1–3.5.4 *and* 7.4.2 *hold, we have*

$$(m_h^n)'(\alpha) \ge \beta_* \Delta t, \tag{7.51}$$

*where constant $\beta_* > 0$ is independent of $h$ and $\Delta t$.*

*Proof.* For a small $\Delta\alpha > 0$, we have

$$\begin{aligned}
m_h^n(\alpha + \Delta\alpha) - m_h^n(\alpha) &= M_h^n(\tilde{R}_h^n(\alpha + \Delta\alpha) \setminus \tilde{R}_h^n(\alpha)) \\
&\ge \tilde{\rho}_* |\tilde{R}_h^n(\alpha + \Delta\alpha) \setminus \tilde{R}_h^n(\alpha)|,
\end{aligned}$$

where the last inequality is obtained by Assumption 7.4.2. By the estimate of volume $|\tilde{R}_h^n(\alpha+\Delta\alpha) \setminus \tilde{R}_h^n(\alpha)|$ in Lemma 3.5.3, we obtain (7.51) and complete the proof. □

Now we can extend Lemma 3.5.2 to compressible flows.

**Lemma 7.4.4** (Ring adjustment for compressible flows)**.** *Let $R \subset \Omega$ be a ring to be adjusted, and assume the volumes*

$$|R| + |\tilde{R}_h^n(0)| \le C'h \tag{7.52}$$

*and the perimeter*

$$|\partial \check{R}^n| \le C' \tag{7.53}$$

*for some constant $C' > 0$. If Assumptions* 3.5.1–3.5.4 *and* 7.4.2 *hold, then there exists some $\alpha^*$ such that*

$$m_h^n(\alpha^*) = M_h^{n+1}(R), \tag{7.54}$$

*where $|\alpha^*| \le C\big(h + (h/\Delta t + 1)\varepsilon_{\text{flow}}\big)$ for some constant $C > 0$ independent of $n$, $h$, and $\Delta t$.*

*Proof.* Define $M^k(S) := \int_S \tilde{\rho}^k \, d\mathbf{x}$ to be the exact mass of bulk fluid in $S$ at time $t^k$. Without wells, the exact local mass conservation of bulk fluid in the ring $R$ gives $M^{n+1}(R) = M^n(\check{R}^n)$.

For any $\alpha$ in a neighborhood of zero, consider the difference

$$\begin{aligned}
m_h^n(\alpha) - M_h^{n+1}(R) &= (m_h^n(\alpha) - m_h^n(0)) + (m_h^n(0) - M^n(\tilde{R}_h^n(0))) \\
&\quad + (M^n(\tilde{R}_h^n(0)) - M^n(\check{R}_h^n)) + (M^n(\check{R}_h^n) - M^n(\check{R}^n)) \\
&\quad + (M^n(\check{R}^n) - M_h^{n+1}(R)) \\
&= (m_h^n)'(\xi)\,\alpha + (M_h^n(\tilde{R}_h^n(0)) - M^n(\tilde{R}_h^n(0))) \\
&\quad + (M^n(\tilde{R}_h^n(0)) - M^n(\check{R}_h^n)) + (M^n(\check{R}_h^n) - M^n(\check{R}^n)) \\
&\quad + (M^{n+1}(R) - M_h^{n+1}(R))
\end{aligned} \tag{7.55}$$

where $\xi = \xi(\alpha)$ comes from the mean value theorem. For the second and last terms on the right hand side, by (7.23) and (7.52),

$$\begin{aligned} &|M_h^n(\tilde{R}_h^n(0)) - M^n(\tilde{R}_h^n(0))| + |M^{n+1}(R) - M_h^{n+1}(R)| \\ &\quad \le \|\tilde{\rho}_h^n - \tilde{\rho}^n\|_{1,\tilde{R}_h^n(0)} + \|\tilde{\rho}^{n+1} - \tilde{\rho}_h^{n+1}\|_{1,R} \\ &\quad \le (|\tilde{R}_h^n(0)| + |R|)\|\tilde{\rho}_h - \tilde{\rho}\|_\infty \le C' h\varepsilon_{\text{flow}}. \end{aligned} \tag{7.56}$$

For the third term on the right hand side of (7.55), by (3.62),

$$|M^n(\tilde{R}_h^n(0)) - M^n(\check{R}_h^n)| \le \tilde{\rho}^* \big|(\tilde{R}_h^n(0) \setminus \check{R}_h^n) \cup (\check{R}_h^n \setminus \tilde{R}_h^n(0))\big| \le C' h\Delta t. \tag{7.57}$$

For the fourth term on the right hand side of (7.55), by (7.53) and Lemma 7.4.2,

$$\begin{aligned} |M^n(\check{R}_h^n) - M^n(\check{R}^n)| &\le \tilde{\rho}^* \big|(\check{R}_h^n \setminus \check{R}^n) \cup (\check{R}^n \setminus \check{R}_h^n)\big| \\ &\le C''|\partial\check{R}^n|\,\varepsilon_{\text{flow}}\Delta t^n \le C''C'\varepsilon_{\text{flow}}\Delta t. \end{aligned} \tag{7.58}$$

Combining (7.55)–(7.58), and Lemma 7.4.3 gives

$$m_h^n(\alpha) - M_h^{n+1}(R) \le C'''(h\varepsilon_{\text{flow}} + h\Delta t + \varepsilon_{\text{flow}}\Delta t) + \beta_*\Delta t\alpha < 0 \tag{7.59}$$

when $\alpha < -C'''\big(h + (h/\Delta t + 1)\varepsilon_{\text{flow}}\big)/\beta_*$, and

$$m_h^n(\alpha) - M_h^{n+1}(R) \ge -C'''(h\varepsilon_{\text{flow}} + h\Delta t + \varepsilon_{\text{flow}}\Delta t) + \beta_*\Delta t\alpha > 0 \tag{7.60}$$

when $\alpha > C'''\big(h + (h/\Delta t + 1)\varepsilon_{\text{flow}}\big)/\beta_*$. By the continuity of $m_h^n(\alpha) - M_h^{n+1}(R)$, inequalities (7.59) and (7.60) imply that there exists some $\alpha^*$ such that equation (7.54) holds, where $|\alpha^*| \le C\big(h + (h/\Delta t + 1)\varepsilon_{\text{flow}}\big)$. □

Similarly, let $\check{E}^n$ and $\check{E}_h^n$ be the exact trace-back regions of element $E$ with velocities $\mathbf{v}$ and $\mathbf{v}_h$, respectively. Let $\tilde{E}_h^n(\alpha^*, s)$ be the trace-back polygonal approximation of $\check{E}^n$ defined in Lemma 3.5.4 (Figure 3.3, right). The following lemma extends Lemma 3.5.5 to compressible flows.

**Lemma 7.4.5** (Element adjustment for compressible flows)**.** *Let* $m_h^n(\alpha^*, s) := M_h^n(\tilde{E}_h^n(\alpha^*, s))$ *be the numerical mass of* $\tilde{E}_h^n(\alpha^*, s)$*. Assume* (3.70)*, no self-intersected polygons are created during the adjustment, the volumes*

$$|E| + |\tilde{E}_h^n(\alpha^*, 0)| \le C' h^2, \tag{7.61}$$

*and the perimeter*

$$|\partial \check{E}^n| \le C' h \tag{7.62}$$

*for some constant* $C' > 0$*. Then there exists some* $s^*$ *such that*

$$m_h^n(\alpha^*, s^*) = M_h^{n+1}(E), \tag{7.63}$$

*where* $|s^*| \le C(h\varepsilon_{\text{flow}} + h\Delta t + \varepsilon_{\text{flow}}\Delta t)$ *for some constant* $C > 0$ *independent of* $n$*,* $h$*, and* $\Delta t$*.*

*Proof.* For any $s$ in a neighborhood of zero, consider the difference

$$\begin{aligned} m_h^n(\alpha^*, s) - M_h^{n+1}(E) = {} & (m_h^n(\alpha^*, s) - m_h^n(\alpha^*, 0)) + (m_h^n(\alpha^*, 0) - M^n(\tilde{E}_h^n(\alpha^*, 0))) \\ & + (M^n(\tilde{E}_h^n(\alpha^*, 0) - M^n(\tilde{E}_h^n(0, 0)) + (M^n(\tilde{E}_h^n(0, 0) - M^n(\check{E}_h^n)) \\ & + (M^n(\check{E}_h^n) - M^n(\check{E}^n)) + (M^n(\check{E}^n) - M_h^{n+1}(E)). \end{aligned} \tag{7.64}$$

For the first term on the right hand side, since no self-intersected polygons are created during the adjustment, $\tilde{E}_h^n(\alpha^*, s)$ is monotone in $s$, so by (3.70),

$$|m_h^n(\alpha^*, s) - m_h^n(\alpha^*, 0)| \ge \frac{1}{2}\tilde{\rho}_* |\tilde{\mathbf{x}}_{\text{l}}^n - \tilde{\mathbf{x}}_{\text{r}}^n| \, |s| \ge \frac{1}{2}\tilde{\rho}_* \lambda_* h |s|. \tag{7.65}$$

For the second and last terms on the right hand side of (7.64), notice that $M^n(\check{E}^n) = M^{n+1}(E)$ by the exact local mass conservation, so by (7.61) and (7.23),

$$\begin{aligned} & |m_h^n(\alpha^*, 0) - M^n(\tilde{E}_h^n(\alpha^*, 0))| + |M^n(\check{E}^n) - M_h^{n+1}(E)| \\ & \quad = |M_h^n(\tilde{E}_h^n(\alpha^*, 0)) - M^n(\tilde{E}_h^n(\alpha^*, 0))| + |M^{n+1}(E) - M_h^{n+1}(E)| \\ & \quad \le \|\tilde{\rho}_h^n - \tilde{\rho}^n\|_{1, \tilde{E}_h^n(\alpha^*, 0)} + \|\tilde{\rho}^{n+1} - \tilde{\rho}_h^{n+1}\|_{1, E} \\ & \quad \le (|\tilde{E}_h^n(\alpha^*, 0)| + |E|)\|\tilde{\rho} - \tilde{\rho}_h\|_\infty \le C' h^2 \varepsilon_{\text{flow}}. \end{aligned} \tag{7.66}$$

For the third term on the right hand side of (7.64), by (3.74) and Lemma 7.4.4, we have

$$\begin{aligned}|M^n(\tilde{E}_h^n(\alpha^*,0)-M^n(\tilde{E}_h^n(0,0))| &\le \tilde{\rho}^*(|\tilde{E}_h^n(\alpha^*)\setminus\tilde{E}_h^n(0)|+|\tilde{E}_h^n(0)\setminus\tilde{E}_h^n(\alpha^*)|)\\ &\le 2\tilde{\rho}^*\big(h_{\tilde{E}_h^n(\alpha^*)}+h_{\tilde{E}_h^n(0)}\big)\|\mathbf{v}_h\|_\infty|\alpha^*|\Delta t\\ &\le C''h(h\varepsilon_{\text{flow}}+h\Delta t+\varepsilon_{\text{flow}}\Delta t). \end{aligned} \tag{7.67}$$

For the fourth term on the right hand side of (7.64), by (3.61), we have

$$\big|M^n(\tilde{E}_h^n(0,0))-M^n(\check{E}_h^n)\big|\le C''\tilde{\rho}^*h^2\Delta t. \tag{7.68}$$

For the fifth term on the right hand side of (7.55), by (7.62) and Lemma 7.4.2,

$$\begin{aligned}|M^n(\check{E}_h^n)-M^n(\check{E}^n)| &\le \tilde{\rho}^*\big|(\check{E}_h^n\setminus\check{E}^n)\cup(\check{E}^n\setminus\check{E}_h^n)\big| \\ &\le C''|\partial\check{E}^n|\,\varepsilon_{\text{flow}}\Delta t^n\le C''C'\varepsilon_{\text{flow}}h\Delta t.\end{aligned} \tag{7.69}$$

Combining (7.64)–(7.69) and Lemma 7.4.3 gives

$$m_h^n(\alpha^*,s)-M_h^{n+1}(E)\le\frac{1}{2}\tilde{\rho}_*\lambda_*hs+C'''h(h\varepsilon_{\text{flow}}+h\Delta t+\varepsilon_{\text{flow}}\Delta t)<0 \tag{7.70}$$

when $s<-2C'''(h\varepsilon_{\text{flow}}+h\Delta t+\varepsilon_{\text{flow}}\Delta t)/(\tilde{\rho}_*\lambda_*)$, and

$$m_h^n(\alpha^*,s)-M_h^{n+1}(E)\ge\frac{1}{2}\tilde{\rho}_*\lambda_*hs-C'''h(h\varepsilon_{\text{flow}}+h\Delta t+\varepsilon_{\text{flow}}\Delta t)>0 \tag{7.71}$$

when $s>2C'''(h\varepsilon_{\text{flow}}+h\Delta t+\varepsilon_{\text{flow}}\Delta t)/(\tilde{\rho}_*\lambda_*)$. By the continuity of $m_h^n(\alpha^*,s)-M_h^{n+1}(E)$, inequalities (7.70) and (7.71) imply that there exists some $s^*$, where $|s^*|\le C(h\varepsilon_{\text{flow}}+h\Delta t+\varepsilon_{\text{flow}}\Delta t)$, such that equation (7.63) holds. □

As stated in Remark 3.5.2, in practice, we only solve characteristics $\check{\mathbf{x}}_h$ numerically with an accuracy of order $r>0$ in time, so an error $\mathcal{O}((\Delta t)^r)$ would also enter the error estimate of the perturbed velocity $\tilde{\mathbf{v}}_h$. Finally, applying the constructions in Lemmas 3.5.1 and 3.5.4 to $\mathbf{v}_h$, combining estimate of $\alpha^*$ in Lemma 7.4.3 and estimate of $s^*$ in Lemma 7.4.5, and interpolating local definitions of $\tilde{\mathbf{v}}_h$, we obtain a global $\tilde{\mathbf{v}}_h$ and complete the proof of Assumption 7.4.1.

*Remark* 7.4.2. The error due to the perturbed velocity $\tilde{\mathbf{v}}_h$ in Assumption 7.4.1 is consistent with the result of incompressible flows in Assumption 3.0.1, where we assume no flow approximation. The only extra term $\mathcal{O}((h/\Delta t + 1)\varepsilon_{\text{flow}})$ in the error estimate of compressible flows is purely due to the flow approximation.

# Chapter 8

# Conclusions and Future Directions

We have extended a fully conservative characteristic method proposed by Arbogast and Huang [4] that treats the transport approximation of advection-diffusion equations. This method, the volume corrected characteristics-mixed method (VCCMM), is locally mass conservative by design and locally volume conservative by the Volume Correction Algorithm. We proved the stability property of the method by considering the scheme from an algebraic perspective. Actually, the stability comes from the local volume conservation of the method, and we illuminated the algebraic structure of the scheme.

The central work of this dissertation was to give a proof of convergence and an error estimate of the VCCMM. By a key idea introduced by Arbogast and Wheeler [5], we considered an inexact tracing of points as an exact tracing with a perturbed velocity. Reasons for the perturbation are that we numerically integrate the tracing, use polygonal approximations of trace-back regions, and adjust trace-back points to obtain the local volume conservation. With these considerations, the numerical solution of the exact equation satisfies a perturbed equation *exactly*. For the convergence proof, we introduced an approximation of the $L^1$-error, which is a technique due to Kuznetsov [40], and use the entropy inequality. We proved an overall $L^1$-error estimate $\mathcal{O}(h/\sqrt{\Delta t} + h + (\Delta t)^r)$, where $r > 0$ is related to the accuracy of the characteristic tracing itself. In most cases, $\mathcal{O}(h/\sqrt{\Delta t})$ is the leading term of the error estimate, which is the accumulation of the $L^2$-projection errors. The rest of the error comes from polygonal approximations, points adjustment, and approximate characteristic tracing. The major difficulty

of the proof was to verify the existence and estimate the error of the perturbed velocity field that satisfies the local volume constraint. The VCCMM avoids the CFL constraint on the time step and obtains a higher convergence rate compared with the Godunov's method. The results of some numerical tests in Chapter 5 matched the results of the theoretical proof.

For the implementation of the VCCMM, our code features a data structure of the polyline class, which gives us the flexibility to apply this method to more general meshes. This is consistent with the nature of VCCMM, a finite volume method, which means it should be only weakly dependent on the geometry of the mesh.

We gave some numerical examples of a quarter five-spot problem to compare VCCMM with the characteristics-mixed method (CMM) and Godunov's method. In numerical experience, large time steps can be taken for VCCMM, and so it produces less numerical diffusion compared with Godunov's method. The VCCMM also gives physically relevant solutions with monotone concentration contours in the field compared with the CMM. This is due to the local volume constraint, since, without sources or sinks, the numerical solution $c_{E,h}^{n+1}$ on a grid element $E$ is actually the average of numerical solutions $c_{h,F}^{n}$, where $F$ intersects the trace-back region $\tilde{E}$. Therefore, the VCCMM does not create non-physical, new local extrema in the solution.

We also extended the VCCMM to problems of compressible flows. Since the density of the fluid may change, we measured mass instead of volume as the conserved quantity. A similar, fully conservative algorithm for compressible flows was developed. Similar stability and convergence analyses were presented. The convergence result is consistent with the incompressible case, where the only extra error is introduced by the flow approximation.

Some tentative future directions and possible improvements of this research are presented in the following.

**General meshes.** Due to the finite volume nature of the VCCMM, we should be able to apply this method to meshes other than rectangular meshes. In addition, as mentioned when describing the polyline class, this data structure, used in our code, gives the potential for us to make an implementation on general polygonal meshes, even unstructured meshes. This is due to the fact that our polyline structure gives the connectivity of grid points and is not restricted to the geometry of the grid element. This generalization would allow us to deal with problems on complex domains, problems with some non-rectangular features that have to be captured, and problems requiring grid refinement.

**Implementations on higher dimensional spaces.** Although we only consider two dimensional space to show the existence and error estimate of the perturbed velocity field, the general idea of the overall convergence proof is not restricted by the space dimension. We would like to develop mature and stable algorithms for three-dimensional spaces based on the two-dimensional ideas of characteristic tracing and point adjustment presented herein. However, the connectivity of elements in higher dimensional spaces is much more complicated, and so special care must be taken especially when one chooses "rings" for trace-back volume adjustment. As Assumption 3.5.4 suggestes, one chooses rings such that they are approximately "perpendicular" to the direction of flow, so a minimal distance of adjustment would effectively change volumes, i.e., so that the volume correction is sensitive to trace-back point adjustment. These issues would be more critical in higher dimensional spaces, and further investigation is needed.

**Numerical results of compressible flows.** We would like to implement the VCCMM for compressible flows and test some numerical experiments. The challenge is, due to the implicit time marching scheme in the flow approximation (7.13)–(7.15), we need to solve a *nonlinear vector* equation (7.20). An efficient nonlinear solver should be used or developed according to the structure of the nonlinear equation. It is highly likely that the velocity fields of compressible flows

may change in time, so it is expected that both the flow approximation and transport approximation could require expensive computation, since the characteristic tracing has to be computed at each time step. The problem might be alleviated for some simple cases, e.g., *slightly* compressible flows.

**Coupled systems.** In this dissertation, we assume that the tracer we are studying is *dilute*, so that its concentration $c$ is sufficiently small and does not change the fluid velocity $\mathbf{u}$ and density $\rho$. However, in general, the mass conservation equation of the bulk fluid (7.1) may couple with the transport equation of the tracer (7.25) if $\mathbf{u} = \mathbf{u}(c)$ or $\rho = \rho(c)$. This will result in a *nonlinear* transport equation for the tracer. Furthermore, for this coupled system, the flow approximation and the transport approximation will have interactions and exchange information. Further study and investigation are needed for a consistent algorithm design and possible techniques for proving convergence.

# Bibliography


[1] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. D. Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, *LAPACK Users' Guide*, Society for Industrial and Applied Mathematics, third ed., 1999.

[2] T. Arbogast, *User's guide to Parssim1: The parallel subsurface simulator, single phase*, Tech. Rep. TICAM Report 98–13, The Center for Subsurface Modeling, Texas Institute for Computational and Applied Mathematics, The University of Texas at Austin, Austin, Texas, May 1998.

[3] T. Arbogast, A. Chilakapati, and M. F. Wheeler, *A characteristic-mixed method for contaminant transport and miscible displacement*, in Computational Methods in Water Resources IX, Vol. 1: Numerical Methods in Water Resources, T. F. Russell et al., eds., Southampton, U.K., 1992, Computational Mechanics Publications, pp. 77–84.

[4] T. Arbogast and C. Huang, *A fully mass and volume conserving implementation of a characteristic method for transport problems*, SIAM J. Sci. Comput., 28 (2006), pp. 2001–2022.

[5] T. Arbogast and M. F. Wheeler, *A characteristics-mixed finite element method for advection dominated transport problems*, SIAM J. Numer. Anal., 32 (1995), pp. 404–424.

[6] ——, *A family of rectangular mixed elements with a continuous flux for second order elliptic problems*, SIAM J. Numer. Anal., 42 (2005), pp. 1914–1931.

[7] T. Arbogast, M. F. Wheeler, and N.-Y. Zhang, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[8] D. Arnold, F. Brezzi, B. Cockburn, and L. Marini, *Unified analysis of discontinuous galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001/02), pp. 1749–1779.

[9] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, New York, 1972.

[10] J. B. Bell, C. N. Dawson, and G. R. Shubin, *An unsplit higher-order godunov scheme for scalar conservation laws in two dimensions*, J. Comp. Physics, 74 (1988), pp. 1–24.

[11] A. Bourgeat, M. Kern, S. Schumacher, and J. Talandier, *The COUPLEX test cases: Nuclear waste disposal simulation*, Computational Geosci., 8 (2004), pp. 83–98.

[12] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[13] F. Brezzi and M. Fortin, *Mixed and hybrid finite element methods*, Springer-Verlag, New York, 1991.

[14] Z. Cao, *Fast uzawa algorithm for generalized saddle point problems*, Appl. Numer. Math., (2003), pp. 157–171.

[15] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *An a priori error analysis of the local discontinuous galerkin method for elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1676–1706.

[16] M. A. Celia, T. F. Russell, I. Herrera, and R. E. Ewing, *An Eulerian-Lagrangian localized adjoint method for the advection-diffusion equation*, Advances in Water Resources, 13 (1990), pp. 187–206.

[17] Z. Chen, *Expanded mixed finite element methods for linear second order elliptic problems I*, Tech. Rep. 1219, Institute for Mathematics and its Applications, University of Minnesota, 1994.

[18] Z. Chen, *Finite Element Methods and Their Applications*, Springer-Verlag, Heidelberg, New York, 2005.

[19] Z. Chen, G. Huan, and Y. Ma, *Computational Methods for Multiphase Flows in Porous Media*, SIAM, Philadelphia, 2006.

[20] A. Chilakapati, *Numerical simulation of reactive flow and transport through the subsurface*, PhD thesis, Rice University, Houston, Texas, 1993.

[21] ——, *A characteristic-conservative model for Darcian advection*, Advances in Water Resources, 22 (1999), pp. 597–609.

[22] R. Courant, E. Isaacson, and M. Rees, *On the solution of nonlinear hyperbolic differential equations by finite differences*, Comm. Pure Appl. Math., 5 (1952), p. 243.

[23] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, Berlin Heidelberg, 2005.

[24] H. K. Dahle, R. E. Ewing, and T. F. Russell, *Eulerian-Lagrangian localized adjoint methods for a nonlinear advection-diffusion equation*, Comput. Methods Appl. Mech. Engrg., 122 (1995), pp. 223–250.

[25] C. Dawson, *Godunov-mixed methods for advection-diffusion equations in multidimensions*, SIAM. J. Numer. Anal., 30 (1993), pp. 1315–1332.

[26] ——, *The Pk+1-Sk local discontinuous Galerkin method for flow problems*, SIAM. J. Numer. Anal., (to appear).

[27] C. Dawson and M. F. Wheeler, *An operator-splitting method for advection-diffusion-reaction problems*, in MAFELAP Proceedings VI, J. A. Whiteman, ed., Academic Press, 1988, pp. 463–482.

[28] C. N. Dawson, T. F. Russell, and M. F. Wheeler, *Some improved error estimates for the modified method of characteristics*, SIAM J. Numer. Anal., 26 (1989), pp. 1487–1512.

[29] M. Dobrowolski, *$L^\infty$-convergence of linear finite element approximation to nonlinear parabolic problems*, SINUM, 17 (1980), pp. 663–674.

[30] J. Douglas, Jr., C.-S. Huang, and F. Pereira, *The modified method of characteristics with adjusted advection*, Numer. Math., 83 (1999), pp. 353–369.

[31] J. Douglas, Jr., F. Pereira, and L.-M. Yeh, *A locally conservative Eulerian-Lagrangian numerical method and its application to nonlinear trans-*

*port in porous media*, Comput. Geosci., 4 (2000), pp. 1–40.

[32] J. Douglas, Jr. and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[33] R. E. Ewing, T. F. Russell, and M. F. Wheeler, *Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics*, Comput. Methods Appl. Mech. Engrg., 47 (1984), pp. 73–92.

[34] J. D. Foley, A. V. Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice*, Addison-Wesley Professional, 1995.

[35] S. M. F. Garcia, *Improved error estimates for mixed finite element approximations for nonlinear parabolic equations: the continuous-time case*, Numer. Methods for Partial Differential Equations, 10 (1994), pp. 129–147.

[36] ——, *Improved error estimates for mixed finite element approximations for nonlinear parabolic equations: the discrete-time case*, Numer. Methods for Partial Differential Equations, 10 (1994), pp. 149–169.

[37] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, 1984.

[38] R. W. Healy and T. F. Russell, *Analytical tracking along streamlines in temporally linear Raviart-Thomas velocity fields*, in Computational Methods in Water Resources XIII, Vol. 2, Bentley et al., eds., Rotterdam, 2000, A. A. Balkema, pp. 631–638.

[39] ——, *Treatment of internal sources in the finite-volume ELLAM*, in Computational Methods in Water Resources XIII, Vol. 2, Bentley et al., eds., Rotterdam, 2000, A. A. Balkema, pp. 619–622.

[40] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first-order quasilinear equation*, USSR Comp. Math. and Math. Phys. 16, 6 (1976), pp. 105–119.

[41] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, second ed., 1992.

[42] B. J. Lucier, *Error bounds for the methods of glimm, godunov and leveque*, SIAM J. Numer. Anal., 22 (1985), pp. 1074–1081.

[43] M. R. Ohm, H. Y. Lee, and J. Y. Shin, *Error estimates for discontinuous Galerkin method for nonlinear parabolic equations*, J. Math. Anal. and Appli., 315 (2006), pp. 132–143.

[44] O. Pironneau, *On the transport-diffusion algorithm and its applications to the navier-stokes equations*, Numer. Math., 38 (1982), pp. 309–332.

[45] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, I. Galligani and E. Magenes, eds., no. 606 in Lecture Notes in Math., Springer-Verlag, New York, 1977, pp. 292–315.

[46] B. Rivière and M. Wheeler, *A discontinuous Galerkin method applied to nonlinear parabolic equations*, in Discontinuous Galerkin methods, Springer, Berlin, 2000, pp. 231–244.

[47] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., vol. 2, Elsevier Science Publishers B.V. (North-Holland), Amsterdam, 1991, ch. Finite Element Methods (Part 1), pp. 523–639.

[48] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in The Mathematics of Reservoir Simulation, R. E. Ewing, ed., no. 1 in Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, 1983, pp. 35–106, Chapter II.

[49] H. Wang, *An optimal-order error estimate for MMOC and MMOCAA schemes for multidimensional advection-reaction equations*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 69–84.

[50] H. Wang and M. Al-Lawatia, *A locally conservative Eulerian-Lagrangian*

*control-volume method for transient advection-diffusion equations*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 577–599.

[51] H. Wang, D. Liang, R. E. Ewing, S. L. Lyons, and G. Qin, *An ELLAM approximation for highly compressible multicomponent flows in porous media. Locally conservative numerical methods for flow in porous media*, Comput. Geosci., 6 (2002), pp. 227–251.

[52] H. Wang, W. Zhao, and R. E. Ewing, *A numerical modeling of multicomponent compressible flows in porous media with multiple wells by an Eulerian-Lagrangian method*, Comput. Vis. Sci., 8 (2005), pp. 69–81.

[53] M. F. Wheeler, *A priori $L_2$ error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal., 10 (1973), pp. 723–759.

# Index

# Vita

Wenhao Wang was born in Xi'an, China on March 4, 1983. He graduated from the Middle School attached to Northwestern Polytechincal University in July 2001. He received a Bachelor of Science degree in Mathematics from Peking University in June 2005. He came to the United State in August 2005 to begin study at the University of Texas at Austin. He received his Master of Science degree in Computational and Applied Mathematics there in July 2007. He began his research on characteristic methods in June 2006.

Permanent address: 5106 North Lamar
Austin, Texas 78751

This dissertation was typeset with LaTeX[†] by the author.

[†]LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.